Roberto Behar Gutiérrez
Mario Yepes Arango

Estadística
Un Enfoque Descriptivo.

Tercera Edición

Santiago de Cali, Colombia, Enero de 2007.

Universidad del Valle

Estadística, Un Enfoque Descriptivo
ISBN 958-670-068-2
© Roberto Behar G. 1996, 2007
Mario Yepes A.

Tel: 572-3334903 – 572- 3212167
FAX 572-3398462
e-mail robehar@univalle.edu.co
robehar@yahoo.com

Talleres Gráficos
De Impresora FERIVA S.A.
Cali, Colombia

Prólogo

El gran mérito de la Estadística como disciplina, es proporcionar las herramientas necesarias para obtener conclusiones sobre una población, a partir de una observación de tan sólo una muestra de la misma. La incertidumbre inherente al proceso de generalización es estudiada y medida con base en la teoría de la probabilidad la cual permite tener la información acerca de la confianza asociada con las conclusiones resultantes de la inferencia realizada.

Existen varias maneras de adquirir el conocimiento de los instrumentos que proporciona la inferencia estadística y la habilidad para su aplicación; una de ellas, la tradicional consiste en estudiar en primer lugar, la teoría de la probabilidad y enseguida estudiar la inferencia estadística propiamente dicha; este es el enfoque que involucran la casi totalidad de los libros que circulan en nuestro mercado.

Una segunda manera de visualizar el proceso de aprendizaje, consiste en el desarrollo de una fase exploratoria de los datos que constituyen una muestra o una población si fuera el caso. En esta fase se trata de definir algunos indicadores de rasgos del conjunto que constituye la muestra y luego de procesar los datos, obtener ideas sobre sus propiedades y posiblemente establecer algunas hipótesis sobre el comportamiento de estos rasgos, o sus relaciones en la población.

En esta fase se produce la maduración de muy buena parte de los conceptos básicos que es necesario estudiar con todo el rigor, no sólo en la etapa de inferencia estadística, sino previamente en el estudio de la teoría de la probabilidad; así por ejemplo se trabaja con la función

empírica de densidad de frecuencia, haciendo cálculos con base en datos; la generalización de este concepto constituye la función de densidad de probabilidad de una variable aleatoria. Análogamente se tratan los conceptos de frecuencias condicionales, de funciones empíricas de densidad conjunta, de independencia estadística, cuya prolongación conceptual al hacer referencia a la población, concluye en lo que representan respectivamente, la probabilidad condicional, las funciones de densidad conjunta de probabilidad y la independencia probabilística entre variables aleatorias.

Con lo anterior no se pretende desconocer que la teoría de la probabilidad puede desarrollarse exclusivamente con base a su estructura axiomática y sin apoyo intuitivo alguno. No obstante, los autores del presente texto, visualizan la teoría de la probabilidad como un instrumento de apoyo que permite el desarrollo de la Estadística para su aplicación; en este sentido, acompañar los tratamientos rigurosos de la probabilidad y la inferencia estadística con una visión intuitiva basada en la manipulación de datos obtenidos de procesos reales, cobra una gran importancia desde el punto de vista de la aplicabilidad de las herramientas teóricas que se estudien. Por tanto esta primera fase objeto de este texto constituye un enfoque descriptivo que enriquece los elementos que permiten interpretaciones intuitivas, que no son un reemplazo del estudio riguroso de las potentes herramientas estadísticas, pero si constituyen un fértil abono para su desarrollo y motivado tratamiento.

Como esta primera fase exploratoria no involucra el tratamiento de la incertidumbre que se genera al inferir, no se requiere del conocimiento de la teoría de la probabilidad, lo cual trae la ventaja adicional de que en caso de no terminar el proceso de estudio completo, la persona que ha experimentado esta fase descriptiva, adquiere elementos importantes para la comunicación de situaciones y problemas en términos estadísticos de tal forma que se le facilita expresar a quien puede asesorarle lo que necesita resolver.

Este texto pretende orientar la primera fase mencionada, por tanto puede ser utilizada por algunos investigadores que deseen hacer acopio de instrumentos de ayuda exploratoria.

Por el contenido, por la metodología y por el nivel de prerrequisitos puede ser usado por todos aquellos estudiantes que vayan a introducirse en la disciplina estadística. En algunos temas se requiere el conocimiento de los rudimentos del cálculo diferencial, aunque no son indispensables para el entendimiento de los conceptos básicos.

En lo que respecta a la metodología para el logro de objetivos planteados, ésta trata en lo posible de mantener la siguiente estructura: en primer lugar el planteamiento de la situación problema que será resuelta por la herramienta que se pretende presentar enseguida; luego se plantea un ejemplo, el cual se utiliza para introducir elementos que permitirán definir la notación simbólica y presentar para el caso concreto del ejemplo, la ilustración de la solución al problema general planteado; por último la presentación general de la herramienta usando la notación definida. Al final de cada capítulo se proponen ejercicios con el objeto de que el lector pueda evaluarse y retomar algunos temas que no hayan quedado suficientemente entendidos.

El contenido del texto es el siguiente: el primer capítulo es una introducción, en la cual se pretende precisar los alcances y la utilidad de la Estadística y ubicar la temática que trata este trabajo, en el contexto de la metodología estadística.

En el segundo capítulo se presenta el tratamiento de los datos provenientes de la observación de una característica en los elementos de una muestra, definiendo algunos rasgos que pueden ser de interés. En el tercer capítulo se hace tratamiento de datos provenientes de la observación de dos características a cada uno de los elementos de una muestra, con el propósito de estudiar su distribución, indicadores de asociación y se desarrolla el concepto de análisis de la varianza. En el cuarto capítulo se trata el modelo de regresión simple, su construcción, su interpretación y sus limitaciones.

Con respecto al uso del texto en el desarrollo formal de un primer curso de Estadística, el docente según los objetivos y de acuerdo con el grupo específico de estudiantes, podrá omitir o no los

desarrollos que impliquen procedimientos matemáticos que no estén al alcance de sus alumnos o no los considere pertinentes, haciendo énfasis en la interpretación de los resultados.

No obstante que este texto es el producto del desarrollo de numerosos cursos, damos excusas por los errores que pudiera presentar y agradecemos las sugerencias o rectificaciones que puedan hacernos con el propósito de mejorarlo con base en la valiosa retroalimentación que debe generar su uso.

Prólogo a la segunda edición

Hoy después de 10 años de uso masivo de esta obra, que ha servido a centenares de estudiantes de las más variadas disciplinas que van desde los propios estudiantes de la carrera de Estadística de la Universidad del Valle, estudiantes de Administración de Empresas, Contaduría, Matemáticas, Ciencias Sociales y Económicas y muchas más, de casi todas las Universidades de la región, estamos entregando a usted, esta segunda edición, en la que se incluyen algunas modificaciones, resultado de las sugerencias de muchos colegas que han visto en el texto un buen instrumento para el logro de sus objetivos.

Se han incluido algunos temas nuevos, se ha profundizado y ampliado el tratamiento de otros, se han aumentado el número de problemas de final de capítulo y se han adicionado explicaciones a algunos tópicos. Conscientes de la gran variedad de disciplinas que son usuarias del texto hemos incluido una gran variedad de referencias bibliográficas.

El gran valor del texto, continúa siendo darle vida a los resultados, no quedarse en las frías cifras, no conformarse con cálculos con base en formulas. Se abunda en interpretación, se enfatiza en los conceptos, que es lo que garantiza en ultimas el desarrollo de criterios para enfrentar futuros problemas y situaciones reales.

Queremos agradecer las valiosas sugerencias de nuestros queridos colegas que durante todos estos años han sido usuarios de esta obra, honrando nuestro esfuerzo, sugerencias que en su mayoría han quedado plasmadas en esta segunda edición. Profesores como: Rafael A. Klinger A.,

Francisco A. Quiroga Z., Jorge E. Delgado, Javier Olaya, Jorge Payán, Robby Nelson Díaz, Hernando Solano H., Guillermo Valdés, Libardo Farfán, Oscar Gamboa, Jaime E. Pérez, Ana María Sanabria, Jorge Rodríguez, Gustavo Vargas, Alexander Taborda, Marco Fidel Suarez, Marco A. Triana, Clara Inés Perea, Antonio Escudero A., Omar Rada B., Huber Ramos, Olga Arias, Viviana Vargas, Mercedes Andrade, William Sánchez, Gabriel Conde, Edwin Rengifo, Heberth Muriel, Reynaldo Carvajal, Hugo Hurtado, Rodrigo Izquierdo, Luis Eduardo Girón, entre muchos otros.

Deseamos agradecer de manera muy particular al ingeniero Jaime Felipe Múnera quien puso todo su profesionalismo y su cariño en el diseño de la nueva edición.

Expresamos nuestro reconocimiento a nuestra querida ex alumna Virginia Cabrera, por la labor de transcripción y edición de este libro, la cual desarrolló no solo con gran profesionalismo sino también con mucho tesón y gran afecto.

Agradecemos a los cientos de alumnos nuestros, muchos de los cuales son ahora profesionales de éxito, quienes compartieron en forma directa la experiencia de ingresar al mundo de la estadística, teniendo en muchas de sus noches este texto como interlocutor y compañero, quienes en su momento nos hicieron notar algunos errores tipográficos, algunos cálculos errados y en no pocas veces sesudas sugerencias.

Roberto Behar y Mario Yepes

Prólogo a la tercera edición

Esta edición, la tercera, resulta de la intención de los autores de hacer público y disponible en la web en forma gratuita, este libro. Por esta razón y para hacer más agradable la lectura hemos ampliado los espacios entre líneas.

Se ha eliminado la fe de erratas, corrigiendo los errores tipográficos, o por lo menos disminuyéndolos.

Otro cambio de interés, Aprovechando las sugerencias de algunos colegas, entre ellos Eloina Mesa y Víctor González, hemos adaptado la notación en lo relativo a la representación de la frecuencia relativa, cambiando la "h" por "f" , induciendo un cambio a la notación de la densidad de frecuencia de h^* hacia f^* y análogamente la frecuencia relativa acumulada de H(x) hacia F(x)..

Estos cambios están más acordes con la notación de la mayoría de los libros, haciendo a los estudiantes más fácil la consulta de otros libros y materiales relacionados así como también empalma de manera más natural con la notación usada en la teoría de la probabilidad para conceptos equivalentes a los aquí desarrollados.

También por sugerencia de algunos colegas que han usado el libro por muchos años, hemos incluido algún desarrollo que ligue el concepto de variable continua en estadística descriptiva con el de variable aleatoria en teoría de la probabilidad, generando un puente intuitivo entre la función de densidad de frecuencia relativa con la función de densidad de probabilidad, relacionando también el área de los rectángulos de un histograma con le área bajo una curva y por supuesto en su definición operativa, las áreas de rectángulos por la integral de la función de

Roberto Behar y Mario Yepes

densidad, haciendo natural el paso de la Función de Distribución Empírica acumulada a su homóloga en probabilidad.

En el capítulo 1, se ha adicionado el apartado "Probabilidad, Estadística y el Método en Ingeniería", que corresponde casi textualmente a un artículo que los profesores del área de estadística de la Escuela de Ingeniería Industrial, publicamos en la revista "Ingeniería y Competitividad" de la facultad de ingeniería de la Universidad del Valle.

Capítulo 1

INTRODUCCION

1.1 HISTORIA DE LA PALABRA ESTADÍSTICA [1]

En su sentido actual, las palabras estadística y estadístico (esta ultima como sustantivo o como adjetivo) tienen menos de un siglo de existencia, pero se emplean desde hace más tiempo, siendo interesante estudiar el proceso por el que han llegado a adquirir la significación que hoy tienen.

[1]Yule-Kendall: "Introducción a la Estadística". Editorial Aguilar. Edición 14. 1967. Pags. 6, 7 y 8.

Las palabras estadista, estadística, estadístico, parece que derivan más o menos indirectamente del latín STATUS, en el sentido adquirido en el latín medioeval, de un estado político.

La primera de las tres palabras citadas es mucho más antigua que las otras dos. La palabra estadista se encuentra, por ejemplo en "Hamlet" (1602), en "Cimbelino" (1610 ó 1611) y en "El paraíso recobrado" (1617).

Según parece, la palabra estadística se empleó por primera vez en "Elementos de erudición universal" del barón J.F. Von Bielfeld, traducido al inglés por W. Hooper M.D. (vol.3, Londres 1770), uno de sus capítulos se titula "**estadística**" y en él se define ésta como "La ciencia que nos enseña la situación política de los estados modernos del mundo conocido". La palabra **estadística** aparece de nuevo con una definición quizás más amplia, en el prefacio de "Una visión política del estado actual de Europa" por E.A.W. Zimmermann publicada en 1787.

"Hace aproximadamente cuarenta años -dice Zimmermann- que esta rama del conocimiento político, que tiene por objeto estudiar la potencia real y relativa de los distintos estados modernos, de la capacidad derivada de sus condiciones naturales, la industria y la civilización de sus habitantes y la sabiduría de sus gobernantes, se ha constituido, principalmente por parte de los escritores alemanes, en una ciencia independiente... por la forma mas conveniente que ahora ha tomado... esta ciencia conocida por el recién inventado nombre de **estadística**, ha llegado a ser un estudio favorito en Alemania" ; y el adjetivo aparece también: "A los diversos artículos contenidos en esta obra, algunos acreditados escritores **estadísticos** han añadido un resumen de las principales épocas de la historia de cada país".

En pocos años estos vocablos fueron aceptados por diversos escritores, especialmente por Sir John Sinclair, el editor y organizador de la primera "Información estadística de Escocia" al cual se ha atribuido frecuentemente su introducción. En la carta circular dirigida al clero de la iglesia de Escocia en mayo de 1790, indica que en Alemania las llamadas "investigaciones estadísticas" han alcanzado gran extensión, y añade una nota explicativa de la frase "investigaciones

estadísticas" ó "investigaciones relativas a la población, a las circunstancias políticas, a la producción de un país y a otros asuntos de interés público". En la "Historia del origen y progreso...", de la obra citada nos dice: "mucha gente se sorprendió al principio de que yo usara las nuevas palabras **estadística** y **estadístico**, porque suponían que nuestra propia lengua podía expresar el mismo sentido, con algún otro término. Pero en el curso de un largo viaje a través de los países del norte de Europa, que hice en 1786, encontré que en Alemania andaban ocupados en una especie de investigación política a la que habían dado el nombre de ESTADÍSTICA y creyendo que una palabra nueva podría llamar más la atención pública, resolví adoptarla y espero que esté ya completamente naturalizada e incorporada a nuestro idioma".

Esta esperanza estaba ciertamente justificada; pero la significación de la palabra sufrió un rápido desarrollo durante el medio siglo siguiente a su introducción.

"**estadística**" (Statistik), en el sentido en que el término fue empleado por los escritores alemanes del siglo XVIII, por Zimmermann y por Sir John Sinclair, significaba simplemente la exposición de las características más notables de un **Estado**, siendo la forma de exposición casi inevitable en aquel tiempo predominantemente verbal. La condición y el carácter definido de los datos numéricos habían sido reconocidos en época algo anterior -especialmente por los escritores ingleses-, pero las cifras fidedignas eran escasas. Sin embargo, después de comenzar el siglo XIX fueron aumentando los datos oficiales; y en consecuencia las antiguas descripciones verbales fueron desplazadas poco a poco por las exposiciones numéricas. La Estadística adquirió casi insensiblemente una significación más estrecha a saber: la exposición de características de un Estado por métodos numéricos. Difícil es fijar la época en que tal palabra adquirió este significado cuantitativo; pero según parece la transición se realizó sólo a medias, aún después de la fundación de la Royal Statistical Society en 1834. Los artículos del primer volumen del journal aparecidos en 1838-39 son en su mayor parte de carácter numérico, pero la declaración oficial no hace referencia alguna al método. "Podemos decir, con palabras del programa de esta sociedad, que Estadística es la investigación de los hechos objeto de cálculos para poner de manifiesto las condiciones y perspectivas de la sociedad". Se reconoce sin embargo, que "el estadista prefiere utilizar cifras y datos numéricos".

Una vez realizado este primer cambio de significación, siguieron otros. La palabra Estadística utilizada primero como el nombre de una ciencia, fue aplicada después para designar las series de cifras sobre las que aquellas operaba y así se habló de estadísticas vitales, estadísticas de beneficencia y otras. La misma palabra se aplicó luego a datos numéricos similares referentes a otras ciencias, como la Antropología y la Meteorología. A fines del siglo XIX hallamos "estadísticas de niños clasificados en listos, medianos y torpes", "estadísticas de caracteres mentales en el hombre" y hasta "un examen estadístico de las características del hexámetro" de Virgilio.

La evolución del significado del adjetivo "estadístico" (statistical) y del nombre "estadístico" (statician) fue naturalmente análoga.

No hace falta multiplicar los ejemplos para hacer ver que la palabra **estadística** no está hoy vinculada en forma principal a las "**cosas del estado**".

1.2 DIMENSION ACTUAL DE LA ESTADÍSTICA

La estadística ha tenido un desarrollo extraordinario, que ha hecho que muchos problemas que antes no tenían una clara solución, hoy la tengan.

Para que podamos hacernos a una idea de la diversidad de campos en los que la Estadística juega un papel importante, se presentan a continuación algunas situaciones.

1. Prueba de una vacuna

Se quiere determinar la efectividad de una vacuna; para ello se diseña un experimento en el cual participa un gran conjunto de niños de cierta edad, los cuales son clasificados al azar en 2 grupos. Al primer grupo se le aplica una vacuna y al segundo grupo no. Se les hace un seguimiento durante un período adecuado de tiempo para comparar la incidencia de la enfermedad problema en cada grupo. ¿Cuál debe ser la diferencia mínima en el número de afectados para aceptar que la vacuna es efectiva?

2. Determinación de la etiología de una enfermedad

Para que una enfermedad se produzca es preciso una combinación adecuada de las condiciones de tres elementos que son: el agente, el ambiente y el huésped. Al proceso constituido por las interrelaciones de estos tres elementos que caracteriza y explica la presencia de la enfermedad, se conoce como "historia natural de la enfermedad". La Epidemiología se dedica en gran parte a la determinación de la historia natural de las enfermedades, ya conociendo ésta, es posible determinar cuál etapa del desarrollo de la enfermedad es más factible de interrumpir para evitar la misma.

No es fácil en la mayoría de los casos, determinar la historia natural de una enfermedad, y en ello la Estadística juega un papel muy importante al proporcionar herramientas para comparar la distribución de la enfermedad en grupos con diversas características socioeconómicas (sexo, edad, condiciones geográficas, raza, hábitos, etc.), con el ánimo de ir acotando las condiciones ambientales y del huésped que conduzcan a la explicación de la historia natural de la enfermedad.

3. Determinación de la dosis de una droga

Para lanzar una nueva droga al mercado, es necesario superar una serie de etapas y pruebas que son mas o menos rigurosas dependiendo de las leyes del país en cuestión. Generalmente el consumo de una droga puede producir efectos colaterales que pueden ser más o menos graves. Por tal razón es necesario diseñar experimentos para determinar niveles de sensibilidad y la dosis adecuada que permita atacar la enfermedad y no producir molestias. (Nótese que estos aspectos varían de persona a persona).

4. Caracterización de la demanda por el servicio de urgencia hospitalaria

La demanda por el servicio de urgencia hospitalaria es variable de mes a mes, de semana a semana, de día a día, e inclusive en horas de un mismo de día.

El conocimiento de dicha distribución es de mucho interés para la determinación de recursos humanos y materiales y para su programación. Un acercamiento a la distribución de la demanda puede conseguirse recolectando información y realizando algunos análisis estadísticos.

5. Fase de planeación

La planeación es en cierta forma "mirar hacia el futuro con los ojos del pasado". En el proceso de planeación se requiere disponer la información cuantitativa y cualitativamente adecuadas para tomar decisiones ahora, que tendrán implicaciones en el futuro. Una empresa debe hacer proyecciones de demanda del artículo que se produce, pues con base en ella, se hará la programación de la producción y todo lo que ella trae consigo.

Dicha demanda puede ser estimada a través de modelos estadísticos de **series de tiempo**.

6. Control de calidad

La calidad con que se produce un artículo es importante para cada industria. Esta constituye un factor básico de competencia en el peor de los casos, por ejemplo en el caso de drogas o alimentos se trata de la integridad e incluso de la vida de las personas. En la práctica es muy costoso y a veces imposible inspeccionar el 100% de la producción o de la materia prima, se puede en estos casos diseñar un plan estadístico de muestreo, y unos instrumentos que permitan tomar decisiones muy confiables sobre la calidad de un lote de producción a partir de la observación de unos pocos artículos, economizando de esta manera dinero y tiempo.

7. Comparación de la eficiencia de dos procesos

Se desea decidir sobre cuál de 2 procedimientos utilizar para la realización de una actividad intermedia en la producción de un artículo, tomando como criterio de eficiencia. Se diseña el experimento y se realizan observaciones durante corto tiempo con base en las cuales se deberá decidir con cierta confiabilidad cuál procedimiento es mejor.

8. Producción agrícola

Se van a sembrar grandes áreas de terreno con papa china, se requiere por tanto diseñar un experimento para determinar entre otras cosas: ¿cuál debe ser la distancia entre plántulas?, ¿cuáles deben ser los niveles de agua y de nutrientes a usar?, ¿hay o no interacción entre la distancia entre las plantas y los niveles de nutrientes? todo ello para conseguir óptima producción.

9. Econometría

Determinación de las principales características socioeconómicas que generan la inflación y cómo influye cada una de ellas, presentado esto a través de un modelo de regresión.

10. Análisis actuarial

Una empresa de seguros de vida, desea determinar cuanto debe cobrar al año por una póliza, según la edad. Para ello, debe realizar un estudio estadístico sobre los riesgos y las frecuencias de muertes por grupos de edad.

El papel de la Probabilidad en Ingeniería.

Cuando hablamos de ingeniería, casi siempre se piensa en matemáticas, y más generalmente en métodos para la modelación, para el análisis y evaluación de situaciones en las que se planea actuar sobre la naturaleza, para transformarla con algún fin, en armonía con el medio ambiente y considerando la optimización de los recursos.

En la formación de ingenieros, la pertinencia de la probabilidad y de la estadística es bastante evidente. Si tomamos como referencia a Koen (1985), en su libro "El método en Ingeniería", nos percataremos que inherente a su esencia, la estrategia del ingeniero, está envuelta en una nube de variabilidad e incertidumbre, en medio de la cual, debe tomar decisiones que lo acerquen a su objetivo, de una manera heurística. Veamos algunas expresiones textuales del mencionado libro, que refuerzan estos planteamientos:

- *"...Por el método de Ingeniería quiero decir la estrategia para causar el mejor cambio posible, con los recursos disponibles, en una situación incierta o pobremente estudiada"*

Aquí queda implícito que el ingeniero debe tomar decisiones con información incompleta, en ambiente de incertidumbre, asumiendo riesgos, pero no de manera aventurera o irresponsable: lo hará con criterio y guiándose por heurísticas, muchas de las cuales tienen como propósito hacerse buenas ideas sobre la magnitud de los riesgos que asume y saber cual es el lado que lo pone conservadoramente cerca de la seguridad.

El mismo autor, dedica el capítulo 3 de su libro a definir algunos heurismos usados por el método de Ingeniería y los divide en 5 categorías, una de las cuales es:

- *"Algunos heurismos que usan los ingenieros para mantener el riesgo dentro de los límites permitidos"*.

Otras expresiones como:

- *"...nunca será posible desarrollar del todo algunos problemas complicados, debido a la incertidumbre inherente al Método de Ingeniería"*.

- "*Dado que el ingeniero tratará de encontrar la mejor respuesta, aún en situaciones relativamente viables para tomar una decisión, es inevitable que exista algún riesgo. Esto desde luego no significa que todos los niveles de riesgo sean aceptables. Como podría esperarse a esta altura de la discusión, lo que es razonable está determinado por heurismos adicionales que controlan el tamaño del riesgo que el ingeniero está dispuesto a tomar"*.

- *"Si el sistema que desea cambiar es complejo y poco entendido; si el cambio deseado es el mejor disponible y si está limitado por la disponibilidad de recursos, entonces usted*

está ante un problema de Ingeniería. Si usted logra el cambio usando el Método de Ingeniería, entonces usted es un ingeniero.".

Basados en Koen (1985), queda claro que el método de ingeniería y la profesión de ingeniero, estarán limitados en su eficiencia y eficacia, si en un sitio privilegiado de su maletín de heurísticas, no tienen algunas que le permitan resolver y decidir en ambientes de riesgo e incertidumbre, que constituyen su condición natural de operación.

En no pocas ocasiones, el ingeniero deberá inferir información de otros situaciones que a su parecer se han producido en circunstancias similares a la de su interés, generándose así posibles errores, cuyo magnitud deberá ser considerada por él, en la toma de decisiones. Por otro lado muchos problemas en ingeniería involucran procesos y fenómenos naturales que presentan variabilidad y aleatoriedad inherentes, haciendo que ellos no puedan ser descritos o caracterizados de manera exacta. Por estas razones los procesos de planeación y de diseño en ingeniería deben tomar en consideración, casi obligatoriamente, estas consideraciones de aleatoriedad y de incertidumbre.

Cuando Koen se refiere a que no todos los niveles de riesgo son aceptables, está sugiriendo que el ingeniero en su responsabilidad, deberá cuantificar el riesgo para decidir con base en un juicio sobre la magnitud de incertidumbre razonable. De esta manera la formulación de decisiones relacionadas con procesos inciertos, requerirán valoraciones del tipo riesgo-beneficio.

¿Cuál es la naturaleza de aquellas heurísticas que le permiten al ingeniero cuantificar el tamaño del riesgo?

¿Cómo obtener una estimación de la magnitud de un efecto de particular importancia en un proyecto, que garantice al ingeniero que actúa hacia el lado de la seguridad en cuanto al riesgo, pero sin perder de vista la racionalidad económica o práctica?

La Probabilidad, la Estadística y el Método de Ingeniería.

La respuesta a los anteriores interrogantes, la tiene la teoría de la probabilidad y la estadística.

En una situación experimental por ejemplo, en la que se pretende valorar la fatiga de cierto material, es casi seguro, que experimentos repetidos bajo condiciones similares no generarán el mismo resultado. ¿ Cual debe ser entonces el valor de la fatiga que debe reportarse, asociado a dicho material, en un proceso de diseño?.

Si el ingeniero se enfrenta al problema del diseño de un canal para aguas de lluvia, ¿cuales deben ser sus parámetros de diseño si el quisiera que el canal fuera suficiente, para lluvias tan intensas como aquellas que se presentan en promedio una vez cada diez años?.

Conociendo la imposibilidad de predecir con certeza de que magnitud serán las máximas lluvias que ocurrirán en el futuro. Cómo responder la pregunta?

El ingeniero debe cuantificar el riesgo y las heurísticas que le permitirán hacerlo, son competencia de la probabilidad y la Estadística.

En electrónica, es posible conocer la fiabilidad de cada una de los elementos de un circuito, como poder, a partir de estas probabilidades individuales, conocer el riesgo de falla del circuito completo como un sistema?.

En este camino, conocer los *elementos básicos de la teoría de la probabilidad*, de tal manera que a partir de la estimación de la probabilidad de ocurrencia eventos simples, pueda obtenerse información sobre el riesgo de ocurrencia de eventos compuestos y complejos, es una necesidad para el ingeniero.

Si con un determinado sistema, es posible resolver el problema con un riesgo r, ¿cuál sería el riesgo si se colocaran n sistemas en paralelo? O combinaciones de serie y paralelo?

En una situación pobremente estudiada, ¿cómo hacer predicciones del riesgo, usando información incompleta?

Si la magnitud de un factor F, es un insumo clave para la solución de un problema de ingeniería, pero solo dispongo de algunos datos sobre F, ¿Cómo puedo estimar la magnitud de F, asumiendo un riesgo de equivocarme en la estimación, definido a priori por el ingeniero?

En esta situación la probabilidad y la estadística pueden apoyar la formación del ingeniero proporcionándole las herramientas adecuadas para la construcción de heurísticas, a través de la llamada *estimación de cantidades, por medio de intervalos de confianza*.

Koen (1985) en su intento por caracterizar el trabajo del ingeniero, expresa cómo el ingeniero inicia su trabajo saliendo de un punto de partida que corresponde a una situación de incertidumbre o pobremente estudiada y que su punto de llegada es incierto. En el camino, deberá ir resolviendo las dificultades y obstáculos y tomando decisiones cuando existan varios caminos alternativos.

¿Cómo poder hacer comparaciones y tomar decisiones ante diversos cursos alternativos de decisión, en un ambiente de incertidumbre?

En esta problemática, la probabilidad y la estadística se constituyen en una verdadera mina, de la cual el ingeniero puede dotarse de las heurísticas apropiadas para enfrentar con muy buenas posibilidades de éxito la situación de comparar alternativas, con información parcial, cuantificando el riesgo de tomar una mala decisión. Este yacimiento de heurísticas, se conoce en estadística como *Contraste de hipótesis*. ó *¿Cómo decidir entre varios posibles cursos de acción en ambiente de incertidumbre?*

Koen plantea de manera muy pedagógica la diferencia entre los dominios de la Ciencia y de la Ingeniería. Uno de los elementos conceptuales que marca esta diferencia, es la restricción en los

recursos disponibles. A diferencia de la ciencia, en la ingeniería no se hace referencia a la solución, sino a una solución.

En ingeniería una buena solución no se puede juzgar, sin el conocimiento de la restricción generada por la disponibilidad de recursos.

En ingeniería puede preferirse una solución que no es la óptima absoluta (utilizando algún criterio de optimalidad), pero que se aproxima bastante bien a los requerimientos, si ésta es mucho más rápida y/o barata que la óptima.

Si la recopilación de la información completa requiere de un periodo de tiempo exagerado o exige una cantidad de recursos muy grande, el ingeniero deberá disponer de heurísticas que le permitan saber cuál es el punto de equilibrio entre la cantidad de recursos a invertir en obtener información y la magnitud del riesgo de equivocarse y sus consecuencias al tomar decisiones con dicha cantidad limitada de información.

La probabilidad y la estadística ofrecen un excelente menú, para que el ingeniero disponga de heurísticas que le permitan cuantificar el monto de recursos que debe asignar a la inversión en información y la manera de decidir con dicha información. Esta carta de navegación, se conoce en Estadística como *estimación del tamaño de muestra* y puede relacionar un tamaño de muestra a seleccionar con el riesgo de equivocarse al decidir con ella en algún sentido.

Por otro lado ante la incertidumbre o el pobre conocimiento de la situación, el ingeniero debe disponer de heurísticas que le permitan en algunas ocasiones hacer ensayos en pequeña escala, para predecir el comportamiento de un sistema, anticiparlo tomando las medidas adecuadas, llenándose de argumentos para favorecer un curso determinado de acción. Este es el caso por ejemplo, de los cilindros de prueba, que son construidos con la mezcla de concreto que el ingeniero piensa usar en una obra y que debe someter al laboratorio para verificar su resistencia. De nuevo, casi con seguridad, los cilindros construidos con la misma mezcla, presentarán variabilidad en los resultados de resistencia medidos en el laboratorio. Con esta información,

deberá tomarse una decisión que será aplicada a las mezclas que con las mismas especificaciones se realicen para construir la obra en cuestión. Conociendo la existencia de la mencionada variabilidad ¿cómo estar seguros de que las mezclas que se produzcan se comportarán de la misma manera que la muestra estudiada?.

¿Cómo realizar estos ensayos? ¿Cómo concluir con base en la información obtenida en los ensayos, si se sabe que esa información parcial, no es reproducible en forma exacta si se repitieran los ensayos?.

¿Cómo puede comparar la resistencia de varios diseños de mezclas?.

En esta situación, un excelente socavón, rico en las mejores fuentes para producir heurísticas, lo constituye el *diseño estadístico de experimentos*, el cual no solo plantea muy buenas guías para la ejecución de los ensayos, para garantizar la validez de las conclusiones que se obtengan, sino que permite controlar el riesgo, definiendo a priori, la magnitud de los riesgos que el ingeniero está dispuesto asumir, en el sentido de tomar decisiones equivocadas. Además incluye relaciones esenciales que conectan los recursos a invertir con la calidad de las decisiones. En todo análisis de un diseño estadístico de experimentos, arrojará información de tipo probabilístico.

Cuando se trata de la valoración del impacto de alguna medida o política gubernamental sobre el medio ambiente, generalmente se compara la situación antes y después de la intervención. ¿Cómo saber si las diferencias observadas no se deben tan sólo al azar, sino que pueden atribuirse a la intervención estudiada?.

Ya se dijo que una condición inherente al trabajo de un ingeniero, y que por tanto caracteriza el Método de Ingeniería, es la restricción en la disponibilidad de recursos. Entre varias heurísticas comparables en su eficiencia, el ingeniero podría escoger aquella que exija menos insumos de información y en general que implique menos recursos.

Proteger los recursos, es una de sus misiones permanentes. En este sentido poder predecir el estado final resultante de un curso de acción tomando en consideración características de su

punto de partida, le permitirá disminuir los riesgos de invertir recursos en rectificaciones por deficientes predicciones.

Un indicador importante de contaminación de las aguas con materia orgánica, es la llamada demanda bioquímica de oxígeno, DBO, cuyo proceso de medición en el laboratorio, puede tardar 20 días. Para agilizar este proceso de medición, sería de mucha utilidad asociar medidas más tempranas de este mismo parámetro, con las que resultarían al final del proceso, midiendo por supuesto el riesgo de cometer errores de cierta magnitud. De hecho, este es el sentido del parámetro DBO5, que representa la medición de la demanda bioquímica de oxígeno a los cinco días.

Algo similar ocurre con la resistencia del concreto, que puede alcanzar su valor máximo a los 28 días.

Estos ejemplos de aplicación, podrían generalizarse a situaciones problema donde se requiere el conocimiento de magnitud de F, para tomar una decisión, pero en lugar de conocer F, se conocen X, Y, Z y W, que son mucho más baratas y prácticas de medir que la propia F, surge la pregunta: ¿Cuáles heurísticas permiten al ingeniero tomar decisiones equivalentes con éstas últimas en lugar de F? Entre las características disponibles (X, Y, Z y W), ¿Cuál es el subconjunto mínimo que se requiere y cual es la calidad de las decisiones que se tomen con base en dicho subconjunto? ¿Cómo predecir el valor F correspondiente a un conjunto de valores específico de las características (X, Y, Z y W)?

En esta problemática, la Estadística vuelve a salir a la palestra, poniendo a disposición del ingeniero, los modelos para predecir la magnitud de una característica mediante el conocimiento de otras, a través de los llamados ***modelos de regresión***, midiendo en todo caso, en términos de probabilidad los riesgos de equivocarse en las predicciones o estimaciones.

Si una de las condiciones del punto de partida del ingeniero es la disponibilidad de información sobre un conjunto de características relacionadas con la situación problema, ¿Cómo explorar esta información, para plantear a partir de ella algunas hipótesis que permitan orientar el próximo curso de acción?

En esta fase la Estadística entrega en las manos del ingeniero, algunas estrategias para hacer útiles sus datos, dándoles sentido en el contexto de su problema a través del llamado *Análisis Exploratorio de Datos*.

En la planeación de la producción, por ejemplo, se requiere estimar la demanda por cierto producto. Si se conoce, el comportamiento aleatorio de la demanda en el pasado, de qué manera puede usarse esta información, para predecir el comportamiento de la demanda del futuro?. ¿Cómo valorar que tan fiable es esta predicción?.¿Cuál es el riesgo de que la demanda real que se presente, sea inferior a un cierto valor crítico D_0?

Cuando el comportamiento futuro de una característica, es un parámetro de diseño para un proyecto, se requiere disponer de Heurísticas que permitan sacar provecho del conocimiento sobre cómo se ha comportado dicha variable en el pasado, para hacer pronósticos y estimar su fiabilidad. En este campo, la probabilidad y la estadística proveen los elementos necesarios a través del llamado *análisis de series de tiempo y pronósticos.*

En campos específicos de la ingeniería, en los cuales una característica inherente a la calidad de un producto es el tiempo que trascurre hasta que el producto falla o la duración del tiempo entre fallas, se requiere conocer algunos parámetros que garanticen a priori, la confiabilidad del producto o servicio o para la definición de políticas de mantenimiento de equipos, para la definición de tiempo de garantía, es muy conveniente disponer de heurísticas para la predicción de la fiabilidad, campo fértil de la Estadística a través de la *Teoría de la Fiabilidad,* que no es otra cosa, que la aplicación de la teoría de la probabilidad a esta situación específica.

Si se trata de controlar y mejorar la calidad de productos o procesos en ambientes de incertidumbre y variabilidad, como es la situación normal en la industria manufacturera y en las empresas de servicios, las heurísticas universalmente usadas corresponden al área de *Métodos estadísticos para el control y el mejoramiento de la calidad.*

Si se quiere abordar la calidad desde el propio diseño del producto, intentando conocer la interacción entre los parámetros de diseño del producto o de la operación de un proceso, con características de preferencias o del ambiente del usuario final, se requiere usar la Estadística a través de los llamados *Métodos estadísticos para el logro de la calidad por diseño.*

Citando una vez más a Koen (1985), al empezar su capítulo 1, dice:

- " El uso del Método de ingeniería, en vez del uso de la razón, es la herencia de la humanidad más equitativamente distribuida. Por Método de Ingeniería quiero decir la estrategia para causar, con los recursos disponibles, el mejor cambio posible en una situación incierta o pobremente estudiada. Por Razón, quiero dar a entender la habilidad para distinguir lo verdadero de lo falso."

Esta distinción, indica que la lógica formal, no será el instrumento, que usará el ingeniero para definir sus cursos de acción y para tomar sus decisiones sobre lo que funciona o no funciona, pues como lo explica el propio Koen en su caracterización de *heurismos*, no se garantiza que la aplicación de un heurismo sea siempre válida. Además heurismos diferentes disponibles en el maletín del ingeniero pueden conducir a resultados contradictorios.

En este estado de cosas ¿Cómo decidir sobre la plausibilidad de una heurística o de alguna estrategia, en ambiente de incertidumbre, si no es la lógica formal la que nos rige?

Esta situación se identifica extraordinariamente con lo que se conoce como *Pensamiento Estadístico*, el cual da pautas y guías para valorar un conjunto de datos, con base en la naturaleza

del proceso que los generó, sin comprometerse con la validez categórica de los mismos. Es decir, que unos datos serán tan buenos como el proceso que les dio origen.

Igualmente cuando se requiere comparar cursos de acción, la Estadística proporciona unas guías, que han de seguirse, y hacen plausibles la conclusiones que se obtengan al aplicar unos procedimientos consistentes con dichas guías, aunque no las garantiza al cien por ciento, siempre ofrece información sobre el riesgo de equivocarse en la magnitud establecida.

El pensamiento estadístico, es una dimensión transversal a toda heurística que intente obtener información o tomar decisiones en ambientes de variabilidad e incertidumbre.

Para finalizar, podemos plantear la pregunta ¿Cómo comparar la eficiencia de varias heurísticas en ambientes de incertidumbre o en situaciones pobremente estudiadas?

Una posible estrategia para lograr este propósito, como ya lo discutimos anteriormente, puede darse con base en *la simulación*, la cual permite a costos relativamente bajos predecir el comportamiento de una heurística, en diferentes ambientes y condiciones de partida. Conociendo comportamientos aproximados de las componentes de un sistema y de sus complejas relaciones, puede hacerse uso de las herramientas que proporciona la simulación para obtener resultados empíricos del comportamiento del sistema completo, pudiéndose evaluar la sensibilidad o robustez a ciertas condiciones y ambientes.

La gran conclusión, es que es prácticamente imposible, ignorar el impacto de la variabilidad y de la incertidumbre, que son rasgos omnipresentes, en el contexto del trabajo de un ingeniero. Es necesario entonces, conocer los fundamentos de la teoría de la probabilidad que nos permita involucrar en los análisis la medición del riesgo.

1.3 VALIDEZ DE UNA INVESTIGACIÓN

Cuando se hace referencia a investigación en este contexto, se entiende de la manera más general, como un proceso de búsqueda de conocimiento, sin cualificar la naturaleza del conocimiento

producido, ni su valor en términos de la trascendencia, puede referirse a un complicado estudio astronómico, a la exploración celular en búsqueda de la explicación de algunos procesos químicos que tienen lugar en el núcleo de la célula, como también a cosas de menos generalidad y trascendencia, como la investigación sobre si vale la pena o no aumentar la dosis de abono a un cierto cultivo, de acuerdo con el incremento en el rendimiento que se observe. Un estudio para conocer la opinión política en una zona y en un tiempo determinados.

Nótese que en esta parte, no se pretende asociar investigación con Estadística. No obstante cuando se quiere juzgar la validez de un proceso generador de conocimiento, en cualquier campo, no necesariamente usando la Estadística, aparecen en forma natural dos elementos a considerar y a juzgar:

1.3.1 El mecanismo de observación y la validez externa.

El mecanismo de generación de los datos básicos, que han de servir de cimientos o de materia prima para la elaboración de información. En este primer elemento, la atención se centra en valorar si el mecanismo o instrumento usado registra confiablemente los rasgos que se pretenden observar o medir en el objeto de estudio. Así pues en el caso del astrónomo, quien pretende registrar sus datos, usando un sofisticado telescopio, para estimar algunas distancias entre cuerpos celestes, la pregunta clave es si las distancias registradas por su aparato corresponden a las verdaderas distancias en la realidad, deberá estar razonablemente seguro que a través de su instrumento, no se producen desviaciones significativas[2] pues de no ser asi, el astrónomo deberá estimar la magnitud de estas desviaciones o deformaciones, con el propósito de construir ajustes que corrijan las deficiencias de su instrumento. Es razonable pensar que si lo que mide el astrónomo no se corresponde con la realidad, sus elaboraciones conceptuales, aunque plausibles,

[2] Significativo, en el contexto de la astronomía y de la problemática específica que se aborda. Esto deberá ser materia de nuevas consideraciones.

no necesariamente conducirán a afirmaciones confiables. El instrumento de observación adopta las variadas formas, desde un aparato físico, como en el caso del astrónomo, hasta una sofisticada encuesta que contiene preguntas sesudamente elaboradas con la pretensión de obtener la materia prima para construir categorías sobre complicados conceptos sociológicos o psicológicos. En esta situación la cuestión seria entre otras[3] : en realidad los ítems que contiene el formulario y la manera de relacionarlos para construir las categorías, detectan lo que se quiere detectar?, miden lo que se quiere medir?, pues de no ser así, aun cuando los razonamientos que se realicen sean válidos, sus conclusiones no son confiables. Cuando una investigación satisface esta dimensión, se dice que tiene ***validez externa***.

1.3.2 La lógica del pensamiento y la validez interna.

Una vez se dispone de las observaciones, obtenidas con un proceso o instrumento que posee validez externa, puede decirse que tenemos materia prima con calidad adecuada, que se tiene un punto de partida, unas condiciones iniciales, a partir de las cuales se elaborara un nuevo producto, se generaran afirmaciones simples o muy complejas sobre el objeto de observación, que constituyen nuevos "hallazgos".

La valoración de ese nuevo producto, de ese cuerpo de afirmaciones, tiene varias aristas. Una de ellas es la compatibilidad con el conjunto de proposiciones aceptadas como validas, en el campo que se trata. Si se encuentran contradicciones, se está frente a un nuevo problema a resolver: o se rechazan las nuevas afirmaciones y se buscan razones que justifiquen su invalidez o se replantean las proposiciones aceptadas y dadas como válidas hasta ese momento, buscando una explicación plausible para ese nuevo comportamiento registrado. La otra arista, no excluyente con la primera, es juzgar el producto, es decir el nuevo conjunto de afirmaciones generadas, con base en un

[3] Entre otras, que mas tarde abordaremos en forma específica, como lo es la representatividad de la muestra objeto de la aplicación del instrumento.

juicio sobre el proceso de elaboración, es decir haciendo una valoración crítica de "la lógica"[4] utilizada, partiendo de las observaciones válidas, y usando el universo de proposiciones aceptadas como válidas.

Cuando el resultado de esta valoración crítica del proceso de construcción de las conclusiones, es positivo se dice que el estudio tiene **validez interna.**

Los conceptos de validez externa y validez interna, adoptan formas muy especiales, cuando la naturaleza de la investigación, hace que la observación se realice con base en muestras de individuos de una población que tiene variabilidad en cuanto a las características objeto de la investigación y por tal razón las conclusiones son obtenidas mediante un proceso inductivo, en el cual están presentes ingredientes como el azar y la incertidumbre.

1.4 LA VALIDEZ EN INVESTIGACIONES QUE USAN MÉTODOS ESTADÍSTICOS

1.4.1 Validez externa y representatividad.

La característica esencial de los estudios que usan métodos estadísticos, radica en la observación con base en muestras probabilísticas[5] y las inferencias de naturaleza probabilística, que permiten asociar a sus conclusiones o hallazgos niveles de confianza, como resultado de la componente de aleatoriedad o azar que involucra.

---

[4] Entiéndase en el mas amplio sentido.

[5] Muestra probabilística, para diferenciarla del muestreo intencional, en el que es el juicio del investigador el que decide sobre los elementos a estudiar y por lo tanto las inferencias no son de naturaleza estadística. En adelante siempre que se haga referencia a muestra o a muestreo, entenderemos muestreo probabilístico.

Se puede ver que en esta situación una componente adicional al instrumento de observación propiamente dicho, es la **representatividad de la muestra.**

Sobre la representatividad de una muestra, se ha especulado mucho y es motivo de serias controversias, algunas de las cuales aun tienen vigencia.

Aquí, el criterio para valorar la representatividad de una muestra, tiene dos dimensiones esenciales: el mecanismo mediante el cual se seleccionan las unidades a incluir en la muestra y el número de elementos a incluir en la misma. En resumen: la forma y la cantidad.

La forma de muestrear, es decir el mecanismo para seleccionar la muestra, debe ser tal que se procure plausiblemente conservar la estructura de las características y las relaciones que se quieren observar, que los alejamientos se deban solamente a la acción del azar. Esta afirmación, a veces se operacionaliza con afirmaciones como: "..Todos las unidades de la población deben tener la misma probabilidad de ser seleccionadas en la muestra" algo así como la democracia en la selección de la muestra. aunque podría funcionar algo más flexible, como: " ..El mecanismo de selección[6] debe ser tal que se conozca la probabilidad que tiene cada unidad de la población de ser incluida en la muestra..", esta segunda afirmación, mas general que la primera, exige conocer los ponderadores o pesos que más tarde, en el análisis deberá darse a cada una de las unidades de la muestra para conservar la mencionada estructura de la población.

De hecho cada uno de los llamados modelos de muestreo[7], tiene asociado el conocimiento de la probabilidad que cada unidad de la población tiene de ser seleccionada, así por ejemplo en

---

[6] Nótese que la representatividad de una muestra, se juzga más que por si misma, por el mecanismo que le dió orígen.

[7] En las llamadas poblaciónes finitas, es decir que la población esta conformada por un número conocido N de unidades.

muestreo aleatorio simple[8], la probabilidad es igual para todos (1/*N*). En **muestreo estratificado**, es decir cuando la población se ha clasificado en estratos de tamaño conocido, por ejemplo por estratos socioeconómicos, conformando la muestra con las unidades que se seleccionan al azar de cada uno de los estratos, aquí la ponderación de una unidad depende del estrato a que pertenece y está dada por la proporción que representa la muestra en ese estrato con respecto al tamaño del estrato. Análogamente en modelos como el muestreo por conglomerados, por ejemplo, la población puede estar agrupada en barrios o colonias o comunas. Aquí se escogen algunos barrios al azar. En los barrios seleccionados, se sacan manzanas al azar y luego de las manzanas escogidas se extraen viviendas (muestreo por conglomerados trietàpico). Aquí las ponderaciones se definen de acuerdo al número de barrios (unidades primarias), número de manzanas (unidades secundarias) y al número de viviendas en cada manzana (unidades terciarias). Existe otros modelos como el muestreo sistemático de intensidad *K*, en el cual se da un ordenamiento a las unidades de la población, se selecciona la primera al azar y a partir de ese, se toma una cada *K* unidades.

Pueden existir mezclas de estos modelos básicos y además otros tipos de muestreo que surgen como resultado de consideraciones de eficiencia o de dificultades prácticas.

En resumen, puede decirse entonces, que el establecimiento de un modelo de muestreo, que tenga asociadas probabilidades conocidas de selección de cada una de la unidades de la población, es garantía de que la muestra es representativa (por su forma).

La otra dimensión de la representatividad está relacionada con el **tamaño de la muestra**, sobre el cual existen un gran número de mitos y falsas creencias que se van transmitiendo por generaciones.

---

[8] Todos en un "costal" y se saca al azar del costal una muestra.

Existe la falsa creencia de que para que la muestra sea representativa debe contener el 10% de las unidades de una población, lo cual se contradice con un sencillo ejemplo: para saber el tipo de sangre de una persona, no es necesario extraerle el 10% de la sangre, basta con una sola gota, puesto que se sabe que todas las gotas de sangre de su cuerpo son del mismo tipo. Aquí se nota como el grado de homogeneidad de las unidades toma un papel importante en la definición del tamaño de la muestra. Podría traerse también el caso de la sabia ama de casa que solo prueba una sola cucharadilla de su rica sopa, para tomar con base en ella la decisión de ponerle o no mas sal, eso si, asegurándose de antemano en garantizar la homogeneidad al menear con maestría por todos los rincones de la olla. El tamaño de la muestra si se relaciona con el tamaño de la población a muestrear, pero la heterogeneidad, es decir la variabilidad de la característica de interés, pesa mucho más en su determinación, a tal punto que en poblaciones muy grandes[9], el tamaño de la población no tiene ninguna importancia, es decir que las fórmulas para el cálculo del tamaño de la muestra no toman en cuenta el tamaño de la población,

En todo caso el criterio que define si una muestra de un tamaño determinado, puede considerarse representativa, tiene relación con el **nivel de precisión requerido.** Puede intuirse que entre mas precisión se exija, más grande se requerirá la muestra.

La precisión de una estimación puede expresarse generalmente a través de dos elementos: el **error tolerable** (δ) y la **confianza** (γ) o confiabilidad. El error tolerable es la diferencia que estamos dispuestos a aceptar entre el verdadero valor poblacional (θ)[10] y el calculado con la muestra ($\hat{\theta}_n$)[11] . La probabilidad de que el error tolerable no sea sobrepasado debe ser mayor o

---

[9] En la teoría se conocen como poblaciones infinitas.

[10] Al verdadero valor poblacional, el cual es una constante se le llama **parámetro**.

[11] A la expresion para calcular este valor con base en la muestra se le conoce como **estadístico** y cuando se usa como instrumento para conocer la magnitud del parametro, se le llama **estimador**

igual que el nivel de confianza (γ). De esta manera la expresión de donde se despeja el tamaño de muestra es :

$$P\left[\left|\theta - \hat{\theta}_n\right| \leq \delta\right] \geq \gamma$$

La relación entre el tamaño ***n*** de la muestra y el tamaño ***N*** de la población, para

Una precisión constante especificada, se muestra en la figura 1.1.

Nótese que el tamaño de muestra crece muy lento aún con grandes incrementos del tamaño de la población, asi por ejemplo para $N = 300$ resulta una muestra de

n=120. Sin embargo si el tamaño de la población se duplicará a 600, la muestra sería de 150. Notese que no se duplica. Es más, si N = 900, el tamaño de muestra será de n = 164. Si la población fuese muy grande, digamos N = 1000000, el tamaño de muestra sería n = 200, el cual es el valor límite (tope), como se percibe en la figura, manteniendo en todos los casos el mismo nivel de precisión requerido.

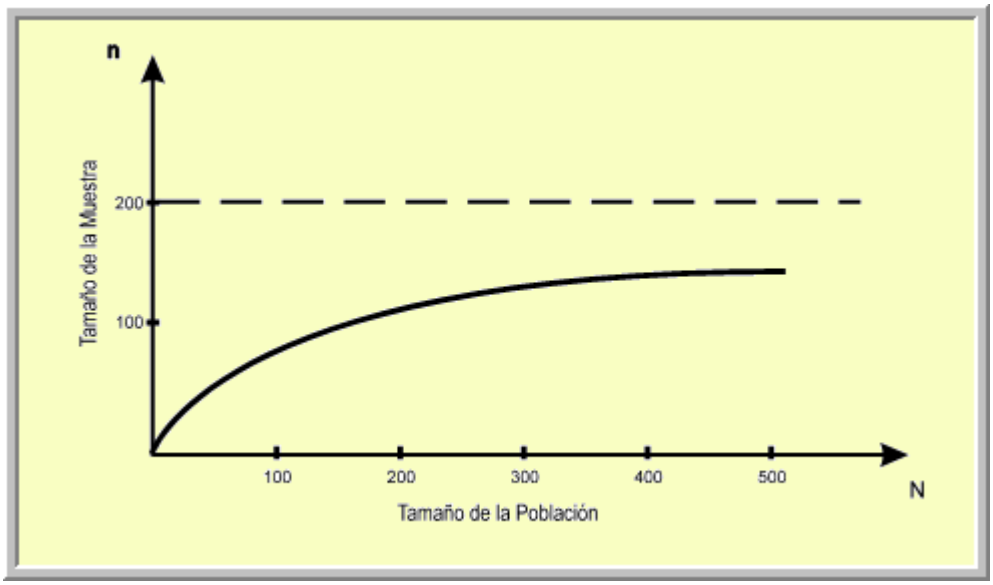


Fig. 1.1. Relación entre el tamaño de la población y el tamaño de una muestra

1.4.2 La validez interna y la comparabilidad.

Cuando en investigaciones que usan la metodología Estadística, se hace referencia a la validez interna, se le esta pidiendo a la lógica de la inferencia estadística, que garantice la **comparabilidad.** Para entender mejor lo esto significa, se presenta una situación donde se viola la comparabilidad: se desea comparar el efecto de la edad de corte de la caña de azúcar, en el rendimiento en toneladas por hectárea, para ello se registra para un buen número de suertes[12] la edad de corte (X) y su rendimiento en Ton/Ha (Y), posteriormente se aplican medidas estadísticas de asociación, para detectar la fuerza de la relación entre estas dos características y resulta una muy pobre asociación, se encuentra posteriormente que las suertes tenían diferente número de cortes[13], lo cual afectaba la comparación, es decir no podría distinguirse si un efecto se debía a la edad o al número de cortes. Un caso extremo podría presentarse si las cañas más jóvenes eran las de mayor número de cortes, pues los dos efectos podrían neutralizarse y hacer aparecer pobre la asociación. En este ejemplo la variable número de cortes, que aparece afectando diferencialmente a las unidades observadas se le conoce como **factor de confusión.**

Podría decirse entonces que la validez interna, la comparabilidad se logra a través del control de los factores de confusión. En esta situación podría encontrarse la asociación de las variables edad de corte y rendimiento, en cada grupo de suertes que tengan el mismo número de cortes, de esta manera, dentro de cada grupo el número de cortes permanece constante y puede lograrse la comparación deseada, siempre y cuando no existan otros posibles factores de confusión, como podrían ser la aplicación de madurantes en forma diferencial en las suertes observadas.

---

[12] Una suerte es un lote de terreno, que se maneja como una unidad, para la siembra, el arreglo, el corte, etc.

[13] Normalmente el terreno se va empobreciendo con el número de siembras (cortes) hasta el punto de que se hace necesario "arreglar" (Remover y abonar) el terreno después de un cierto número de cortes, generalmente cuatro(4).

A esta solución, para lograr validez interna, se le llama construcción de **bloques**[14]. No obstante existen otras soluciones para este mismo problema de falta de comparabilidad, como por ejemplo, la **aleatorización** o involucrar en el modelo de análisis al factor de confusión como una variable, que permite hacer las comparaciones para cada nivel del factor, cuando se da este caso, al factor de confusión en el modelo se le conoce como **covariable.**

Nótese que la identificación de potenciales factores de confusión, no es tarea de un estadístico, sino del investigador que conoce el campo de su disciplina específica.

1.5 ESTADÍSTICA Y MEDICION

La materia prima de la Estadística son los datos, los cuales son el resultado de la "observación" de alguna(s) característica(s) de los elementos de interés en cierto estudio. La naturaleza de la característica y el instrumento que dispone para registrar la misma, definirá el tipo de **escala de medición** que se ajuste a la situación dada.

Escalas de medición. Cuando se hace referencia a las escalas se trata de asociar números a las características con el propósito de manipularlas y obtener nuevo conocimiento sobre las características del estudio.

Se consideran generalmente cuatro escalas de medición: escala nominal, escala ordinal, escala de intervalo y escala de razón.

La escala nominal, hace uso de los números para dar nombre a los elementos que han sido clasificados en distintos grupos, clases o categorías de acuerdo con alguna propiedad cualitativa. El número asignado a una clase sólo actúa como un rótulo o código para diferenciar los elementos de esa clase con los de otra. Por ejemplo si se clasifica un conjunto de objetos por su

[14] De allí el famoso nombre de diseño de bloques al azar

color, las categorías pueden ser: azul, amarillo, rojo, verde, a las cuales podemos asociar respectivamente los números 1,2,3,4 y se hablará de la categoría 1 para hacer referencia al grupo de objetos de color azul o 4 para el verde, pero los números aquí, sólo son códigos para nombrar los elementos de una clase.

La escala ordinal, hace uso de los números para clasificar los elementos de un conjunto en categorías en los cuales los números no sólo sirven para nombrar sino que son base para comparaciones de la forma: "más grande", "igual", "menor", es decir, que el valor numérico de la medida se usa para indicar el orden que ocupa un elemento al comparar el tamaño relativo de sus medidas, del más grande al más pequeño, de allí el nombre de escala. Un ejemplo, cuando a una persona se le pide ordenar de la más importante a la menos importante, asignando números de 1 a 4, a las siguientes necesidades: empleo, salud, vivienda, servicios públicos. Aquí el número se usa para representar la prioridad de las necesidades; de esta manera si un individuo asigna el número 1 a la vivienda y el 4 al empleo, indicará que para él es "más importante" la vivienda que el empleo.

La escala de intervalo, considera pertinente información no sólo sobre el orden relativo de las necesidades, como en la escala ordinal, sino también del tamaño del intervalo entre mediciones, esto es, el tamaño de la diferencia (resta) entre dos medidas. La escala de intervalo involucra el concepto de una unidad de distancia. Por ejemplo la escala con la cual casualmente representamos la temperatura; un incremento en una unidad (grado) de la temperatura está definido por cambio particular en el volumen de mercurio en el interior del termómetro, de esta manera, la diferencia entre dos temperaturas puede ser medida en unidades (grados). El valor numérico de una temperatura es meramente una comparación con un punto arbitrario llamado "**cero grados**". La escala de intervalo requiere un punto cero, como también, una unidad de distancia, pero no importa cual punto se define como cero ni cual unidad es la unidad de distancia. La temperatura ha sido medida adecuadamente por mucho tiempo en las escalas Fahrenheit y centígrada, las cuales tienen diferente temperatura cero y diferentes definiciones de 1 grado o unidad. El principio de la medida de intervalo no es violado por cambios en la escala o en la localización.

La escala de razón, es usada cuando no solamente el orden y el tamaño del intervalo ente medidas son importantes, sino también la razón (o cociente) entre dos medidas. Si es razonable hablar de que una cantidad es "dos veces" otra cantidad, entonces la escala de razón es apropiada para la medición, como cuando medimos distancias, pesos, alturas, etc. Realmente la única diferencia entre la escala de razón y la escala de intervalo, es que la escala de razón tiene un punto cero natural, mientras que en la escala de intervalo éste es arbitrario. En ambas escalas la unidad de distancia es arbitrariamente definida.

Es muy importante tener presente la escala de medición cuando se realiza un estudio, puesto que las pruebas estadísticas varían dependiendo de la escala de medición de las características en referencia.

En general puede decirse que la escala de razón es la que tiene a su disposición una mayor cantidad de herramientas estadísticas para su tratamiento.

1.5.1 Variables discretas y variables continuas.

En las escalas de intervalo y de razón algunas veces es necesario establecer la diferenciación de las variables por su naturaleza, entonces se habla de variables discretas y variables continuas.

Variable discreta, es aquella cuya naturaleza hace que el conjunto de valores que puede tomar la variable sea finito o infinito numerable.

Por ejemplo, la variable: número de personas por hogar, el conjunto de valores que puede asumir ésta son:

$$\{1, 2, 3, 4, \ldots, M\} \quad \text{donde } M \text{ es finito}$$

Otros ejemplos son los siguientes: número de consultas al médico durante un año, número de clientes que llegan a un banco durante una hora, número de ensayos realizados hasta obtener el primer éxito.

Variable continua, es aquella, cuya naturaleza hace que exista un intervalo de puntos, los cuales son valores que puede tomar la variable. Por ejemplo, la estatura de una persona, esta variable puede tomar cualquier valor en el intervalo (1.50 m, 1.60m). El tiempo entre dos llegadas consecutivas al servicio de urgencias de un hospital. El área cultivada de trigo en las fincas del valle del Río Cauca .

Esta clasificación no tiene en cuenta la población en la cual va a ser observada la variable, es decir, no interesa en la clasificación, si la población es finita o infinita, puesto que de acuerdo con la definición una variable es discreta o continua por si misma. Tampoco juega papel alguno el instrumento de medición que se use.

Las definiciones como son presentadas son de utilidad en el tratamiento descriptivo de los datos, como se verá más adelante.

1.6 ALGUNOS TERMINOS USADOS EN ESTADÍSTICA

Se definen a continuación algunos términos que se usarán con frecuencia en el presente escrito.

1.6.1 Población

Se identificará con este nombre al conjunto de elementos de interés en un estudio, sobre los cuales se desea información y hacia los cuales se extenderán las conclusiones. El término población no debe asociarse exclusivamente con población humana; tiene sentido hablar de la población de tornillos que se producen durante un día en una determinada fábrica, o de la población constituida por todas las fincas de un país o una región.

En todo estudio, la población debe estar definida en forma muy precisa, de tal manera que pueda determinarse en algún momento si un elemento dado pertenece o no a la población. Por ejemplo supóngase que se va a realizar un estudio para determinar el porcentaje de desempleo en Cali a abril 4 de 1995; algunas reflexiones tendientes a caracterizar a la población que concierne a dicho estudio son las siguientes:

¿El estudio hace referencia a los caleños o a los residentes en Cali?.

¿Que significa ser residente en Cali? ¿una persona que llegó a Cali en abril 3 de 1995, pertenece a la población? o ¿una persona que se fue de Cali en la misma fecha?

Por la naturaleza del estudio los elementos de interés son las personas que "deberían estar empleadas" (de la observación de estas se definirá quienes lo están y quienes no, para determinar el porcentaje de desempleo), entonces cabe la pregunta: ¿cómo se caracterizan los que "deberían estar empleados"? (edad, condiciones de salud, incapacidad, etc.).

Estas reflexiones sugieren definiciones precisas que conducen a una determinación adecuada de la población.

1.6.2 Muestra

En muchas ocasiones se requiere conocer una característica medible de la población, para ello se puede observar, uno a uno, todos los elementos de la población (Censo), lo cual casi siempre es impracticable o muy costoso; en estos casos puede "hacerse una idea" sobre la característica poblacional, observando sólo algunos elementos de la población, éstos constituyen una muestra de esa población.

1.6.3 Parámetro

Se llamará parámetro a una característica medible **de la población**. Por ejemplo, la edad promedio de los estudiantes de una escuela, el porcentaje de varones; el diámetro promedio de los tornillos que se producen en una fábrica, la tasa de crecimiento promedio de la tilapia roja, el tiempo promedio entre fallas de una maquina etc. Un parámetro es una constante para la población.

1.6.4 Estadística

Se denominará estadística a una característica medible **en la muestra** por ejemplo la edad promedio de una muestra de estudiantes de una escuela, o el porcentaje de varones en la muestra;

el diámetro promedio de los tornillos de una muestra de la población de una fábrica, etc. En general una estadística es una función de los datos de una muestra; como puede intuirse el valor que asume una estadística depende de la muestra que se haya tomado. Generalmente se usan las estadísticas para hacerse una idea de los parámetros, cuando esto sucede se llaman estimadores.

Nótese que una estadística en general varia de una muestra a otra, en este sentido puede mirarse como una variable y dársele el tratamiento que expondremos para las variables.

1.7 ETAPAS DE LA METODOLOGIA ESTADÍSTICA

A continuación se presentan las principales actividades que es necesario realizar cuando se hace un estudio estadístico.

1.7.1. Definición del problema

Consiste en la justificación del estudio, la determinación de los objetivos del estudio, revisión bibliográfica, planteamiento de las hipótesis que se desea probar o rechazar o definición de los parámetros que se desea estimar, incluyendo la precisión que se requiere en la estimación.

1.7.2. Definición de la población

Definir en forma precisa cuál es la población de interés en el estudio, en el sentido presentado en 1.4.

1.7.3. Definición de la estrategia de Análisis

En esta etapa se realiza el plan de análisis, se define una ruta preliminar de ataque al problema. Se seleccionan, si es del caso, algunas técnicas estadísticas que podrían ayudar a esclarecer preliminarmente la situación. Es razonable, que el plan preliminar sufra modificaciones, en la, medida en que se van valorando los hallazgos. Sin embargo tener un plan permite definir un camino de acción, una valiosa guía de acción.

1.7.4. Determinación de las variables de interés

Consiste en la definición de las características de la población que proporcionan la información necesaria para el logro de los objetivos del estudio.

1.7.5. Diseño del estudio

Algunos llaman a esta etapa "diseño del experimento" (o diseño de la muestra) y consiste en definir si se observará la población completa (censo) o sólo parte de ella (muestreo). En este último caso deberá determinarse el tipo de muestreo a utilizar y el tamaño de la muestra para unas especificaciones de precisión deseadas (error tolerable y nivel de confianza), igualmente debe definirse la logística de la recolección de la información.

1.7.6. Recolección de la información

Esta es una etapa muy importante, pues de ella depende la calidad de la información. Los errores en este sentido no los miden las herramientas estadísticas, por esta razón la recolección de la información requiere mucho control sobre los instrumentos como también sobre el proceso de medición.

La dificultad para diseñar un control eficiente sobre la calidad de los datos recogidos, en algunas ocasiones, hace más confiable una muestra que un censo, puesto que se requiere controlar un menor volumen de recursos, garantizando de esta manera una mejor calidad de los datos.

1.7.7. Procesamiento descriptivo de los datos

Esta etapa la constituye la aplicación de las técnicas que proporciona la **estadística descriptiva** y que consiste en la organización de la información en forma útil y comprensible, mediante la elaboración de cuadros, tablas, gráficos y reduciendo los datos recolectados por medio de algunos indicadores que faciliten su interpretación; esta etapa es una fase exploratoria, no obstante constituye un medio para hacerse una idea de los rasgos poblacionales. El análisis de la muestra,

pocas veces tiene interés en sí mismo, siempre se usa la muestra como un instrumento para conocer la población. Por esa razón la característica de Representatividad de la muestra debe garantizarse siempre, independientemente de que se realice análisis exploratorio (descriptivo) o se utilicen herramientas probabilísticas para hacer inferencia estadística.

1.7.8. Inferencia estadística

Se denomina así, al proceso inductivo que permite inferir a toda la población proposiciones, basadas en las observaciones y resultados proporcionados por una muestra. Como puede intuirse en este proceso de inferencia, aparece un factor de incertidumbre, y de error, puesto que muestras distintas pueden arrojar resultados distintos; es precisamente esto lo que hace que la teoría de la probabilidad sea la herramienta básica de la inferencia estadística, ésta no evita los errores que por azar se cometen, pero si los cuantifica y les asocia una medida que indica el nivel de confianza de los resultados obtenidos, lo cual constituye su principal mérito.

1.7.9. Conclusiones y planteamientos de nuevas hipótesis

En esta última etapa se plantean las conclusiones en forma clara, indicando sus alcances y limitaciones, igualmente se plantean nuevas hipótesis que pudieran surgir en la propia exploración de los datos.

1.8 ESTADÍSTICA DESCRIPTIVA

Cuando se habla de estadística descriptiva, da la impresión que es una de las varias "estadísticas" que existen. En realidad es una etapa de la metodología estadística, en la que no se involucra la teoría de la probabilidad como herramienta para realizar inferencias a toda la población, sin embargo se construyen indicadores, se hacen gráficos, se realizan comparaciones, siempre con el interés de conocer sobre la población de donde fue tomada la muestra.

La estadística descriptiva permite procesar los datos de una muestra y obtener información que puede ser usada con fines exploratorios, para plantear hipótesis o como materia prima de la etapa de inferencia estadística.

La complejidad de las herramientas y el volumen de información que se obtenga de una muestra, depende entre otros factores, del número de características que se observen.

En el próximo capítulo se tratará la situación correspondiente a la observación de sólo una variable y se hará referencia a ella como unidimensional.

En los capítulos 3 y 4 se desarrolla la situación en que se observan en la muestra dos variables y se hace mención a ella como bidimensional.

Capítulo 2

Distribuciones Unidimensionales de Frecuencia

2.1 CASO DE UNA VARIABLE DISCRETA

Para considerar este caso, se introduce el siguiente ejemplo:

Ejemplo 2.1

Se toma información sobre el número de clientes que llegan a un banco en una hora pico, observando una muestra de 25 períodos de un minuto se obtuvieron los siguientes resultados: 8, 6, 7, 9, 8, 7, 8, 10, 4, 10, 8, 7, 9, 8, 7, 6, 5, 10, 7, 8, 5, 6, 8, 10, 11.

A esta información, que no ha tenido ningún tipo de tratamiento se le llama **muestra bruta** y se representa por $x_1, x_2, \ldots, x_n$ donde n es el número total de datos.

Se puede comenzar a organizar la información escribiendo los datos distintos de que consta la muestra y haciendo un conteo para determinar el número de veces que aparece cada dato; valor éste que se denominará frecuencia absoluta. El cuadro 2.1 muestra la situación del ejemplo.

CUADRO 2.1

Valor observado (x_i)	Conteo	Frecuencia absoluta (n_i)
4	/	1
5	//	2
6	///	3
7	~~////~~	5
8	~~////~~ //	7
9	//	2
10	////	4
11	/	1
		25

Como puede observarse, la suma de las frecuencias absolutas de todos los datos, debe coincidir con el número total de datos (tamaño de la muestra).

No obstante que la muestra consta de 25 datos, sólo hay 8 datos distintos: 4, 5, 6, 7, 8, 9, 10, 11 que es posible representarlos, sin pérdida de generalidad, como $x_1, x_2,..., x_m$. En nuestro caso $n = 25$ y $m = 8$, de esta manera la **frecuencia absoluta del dato x_i**, se denotará por n_i, así por ejemplo el dato $x_3 = 6$ aparece 3 veces en la muestra, por tanto $n_3 = 3$.

Se puede también expresar la frecuencia absoluta como una fracción o porcentaje del número de datos y surge así lo que se conoce como **frecuencia relativa del dato x_i** que se denota por f_i, así pues:

$$f_i = \frac{n_i}{n} \quad ; \text{ en el ejemplo } \quad f_3 = \frac{3}{25} = 0.12$$

que indica que el dato $x_3 = 6$ representa el 12% de toda la muestra, es decir que de acuerdo con la muestra, en la hora pico, el 12% de las veces llegan al banco 6 clientes por minuto.

También se podría calcular el número de datos que son menores o iguales que x_i, que se denomina **frecuencia absoluta acumulada hasta x_i ,** y se denota por N_i; si $x_1, x_2, \ldots, x_m$ están ordenadas en forma creciente, entonces:

$$N_i = n_1 + n_2 + \ldots + n_i$$

En nuestro ejemplo N_4 es el número de datos que son menores o iguales que $x_4 = 7$, es decir, $N_4 = 11$.

Si la frecuencia absoluta acumulada se expresa como una fracción o porcentaje de toda la muestra, aparece lo que se conoce como **frecuencia relativa acumulada** que se representa por F_i, de esta manera:

$$F_i = \frac{N_i}{n} = f_1 + f_2 + \ldots + f_i$$

Los conceptos, para nuestro ejemplo se sintetizan en el siguiente cuadro de frecuencias.

CUADRO 2.2

CUADRO DE FRECUENCIAS DEL NUMERO DE CLIENTES QUE LLEGAN A UN BANCO EN UN MINUTO DE LA HORA PICO

Valor observado x_i	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia acumulada absoluta N_i	Frecuencia acumulada relativa F_i
$x_1 = 4$	1	0.04	1	0.04
$x_2 = 5$	2	0.08	3	0.12
$x_3 = 6$	3	0.12	6	0.24
$x_4 = 7$	5	0.20	11	0.44
$x_5 = 8$	7	0.28	18	0.72
$x_6 = 9$	2	0.08	20	0.80
$x_7 = 10$	4	0.16	24	0.96
$x_8 = 11$	1	0.04	25	1
Total	25	1.00		

Un resumen de las principales propiedades de las frecuencias se presenta a continuación.

Propiedades y relaciones

Si se toma una muestra de *n* datos, de los cuales hay *m* distintos, que ordenados en forma creciente son $x_1, x_2, \ldots, x_m$, entonces:

- $0 \leq n_i \leq n \; ; \quad i = 1, 2, 3, \ldots, m$
- $n_1 + n_2 + \ldots + n_m = n$; es decir $\sum_{i=1}^{m} n_i = n$
- $f_i = \frac{n_i}{n}$; $0 \leq f_i \leq 1$
- $f_1 + f_2 + \ldots + f_m = 1$; es decir $\sum_{i=1}^{m} f_i = 1$
- $N_j = n_1 + n_2 + \ldots + n_j$; es decir $N_j = \sum_{i=1}^{j} n_i$
- $N_m = n$
- $n_1 = N_1 \leq N_2 \leq \ldots \leq N_m = n$
- $F_j = f_1 + f_2 + \ldots + f_j$; es decir $F_j = \sum_{i=1}^{j} f_i$
- $f_1 = F_1 \leq F_2 \leq \ldots \leq F_m = 1$

En realidad las frecuencias acumuladas pueden definirse como funciones sobre todos los números reales, así:

$N(x)$ = número de datos que son menores o iguales que x

$F(x)$ = fracción (o porcentaje) de los datos que son menores o iguales que x.

Así pues :

$F(4.32)$ = la fracción del total de datos que son menores o iguales que 4.32.

= 0.04

$N(4.32)$ = 1

Para el ejemplo planteado, la distribución $N(x)$, es:

$$N(x) = \begin{cases} 0 & \text{si } x < 4 \\ 1 & \text{si } 4 \le x < 5 \\ 3 & \text{si } 5 \le x < 6 \\ 6 & \text{si } 6 \le x < 7 \\ 11 & \text{si } 7 \le x < 8 \\ 18 & \text{si } 8 \le x < 9 \\ 20 & \text{si } 9 \le x < 10 \\ 24 & \text{si } 10 \le x < 11 \\ 25 & \text{si } x \ge 11 \end{cases}$$

La función $F(x)$ es conocida como **función empírica de distribución acumulativa**, para señalar que ha sido obtenida con base en una muestra de la población, pretendiendo con ella lograr un conocimiento aproximado de la distribución acumulativa que tendría la población (función de distribución acumulativa de probabilidad). A continuación se presenta $F(x)$ para el ejemplo.

$$F(x) = \begin{cases} 0 & \text{si } x < 4 \\ 0.04 & \text{si } 4 \le x < 5 \\ 0.12 & \text{si } 5 \le x < 6 \\ 0.24 & \text{si } 6 \le x < 7 \\ 0.44 & \text{si } 7 \le x < 8 \\ 0.72 & \text{si } 8 \le x < 9 \\ 0.80 & \text{si } 9 \le x < 10 \\ 0.96 & \text{si } 10 \le x < 11 \\ 1 & \text{si } x \ge 11 \end{cases}$$

En general las funciones *N(x)* y *F(x)* pueden definirse de esta manera:

$$N(x) = \begin{cases} 0 & \text{si } x < x_1 \\ N_i & \text{si } x_i \le \mathbf{x} < x_{i+1} \quad j = 1, 2, \ldots, m \\ \mathbf{n} & \text{si } x \ge x_m \end{cases}$$

Análogamente la función empírica de distribución acumulativa

$$F(x) = \begin{cases} 0 & \text{si } x < x_1 \\ F_i & \text{si } x_i \le x < x_{i+1} \quad j = 1, 2, \ldots, m \\ 1 & \text{si } x \ge x_m \end{cases}$$

Las funciones *N(x) , F(x)* son monotónicas no decrecientes, es decir que

si $x_1 < x_2 \Rightarrow N(x_1) \le N(x_2)$ y $F(x_1) \le F(x_2)$.

REPRESENTACIÓN GRÁFICA

Cuando se trate de frecuencias absolutas o de frecuencias relativas, se realizará la representación por medio del llamado **diagrama de frecuencia**, que consiste en colocar en el eje horizontal los valores x_i, que toma la variable y levantando en cada punto un segmento vertical de longitud igual a la frecuencia correspondiente.

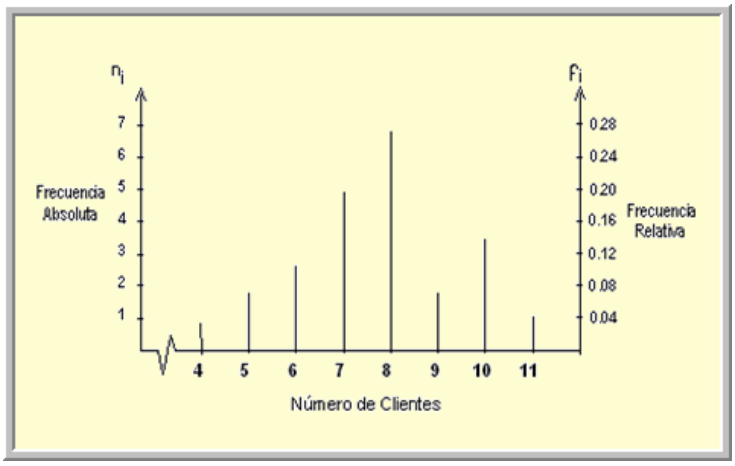


Fig. 2.1. Diagrama de frecuencias del número de clientes que llegan a un banco en un minuto, en la hora pico.

El gráfico de frecuencias absolutas difiere del gráfico de frecuencias relativas sólo en la escala del eje de las ordenadas, por tal razón aparece un solo gráfico con dos ejes: en el eje de la izquierda se leen las frecuencias absolutas y en el de la derecha se leen las relativas.

Cuando consideramos las frecuencias acumuladas, la representación gráfica consiste en llevar a un plano cartesiano las funciones *N(x)* y *F(x)*. Como se aprecia en la Figura 2.2.

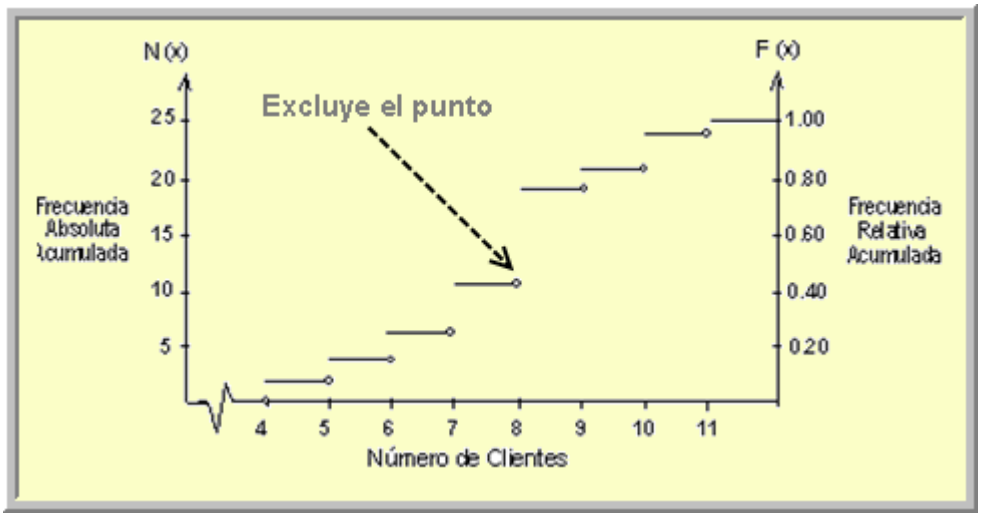


Fig. 2.2. *Gráfico de frecuencias acumuladas para la variable "número de clientes que llegan a un banco en un minuto en la hora pico"*

Como puede notarse el gráfico corresponde a una función escalonada, lo cual indica que sólo hay datos en los puntos de discontinuidad, cuya frecuencia está representada por el valor del salto correspondiente.

2.2 CASO DE UNA VARIABLE CONTINUA

Supóngase que se tienen observaciones sobre la estatura de las personas que conforman una muestra de tamaño 25 y que el instrumento de medición usado tiene precisión hasta las centésimas de milímetro, así pues un valor podría ser 1.74325 metros; si se pretendiera aplicar el procedimiento que se usó para las variables discretas, habría varios problemas, uno de ellos es que seguramente, todos los datos son distintos, lo cual generaría una tabla de frecuencias absolutas con el mismo nivel de información que la muestra bruta; además, no es de interés conocer con ese nivel de detalle la información, por ejemplo, no es de interés conocer cuántas personas tienen una estatura de 1.74325 metros.

En estos casos, es más fácil agrupar la información en los llamados **intervalos de clase.** Para ilustrar sobre su construcción, se plantea el siguiente ejemplo.

Ejemplo 2.2

Los datos que a continuación se presentan corresponden a los tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital:

13.1, 7.1, 14.8, 19.0, 10.2, 18.0, 19.8, 15.0, 17.3, 10.8, 22.3, 14.5, 17.1, 14.9, 12.0, 14.0, 18.4, 10.2, 15.8, 16.5, 15.0, 17.6, 4.2, 13.4, 21.2, 14.7, 13.8, 21.0, 14.3, 11.1, 18.9, 8.3, 16.6, 11.2, 20.2, 14.4, 13.5, 18.2, 12.4, 17.0, 26.7, 15.5, 22.0, 12.9, 17.9, 7.4, 18.0, 19.8, 16.0, 21.2.

Generalmente se empieza por determinar las observaciones extremas (mínima y máxima), que en el ejemplo aparecen marcadas: $\min(x_i) = 4.2$; $\max(x_i) = 26.7$.

Estos valores extremos definen el rango de la muestra:

$$\text{rango} = \max(x_i) - \min(x_i)$$

Se debe determinar los valores $L_0, L_1, L_2, ..., L_m$ que constituirán los límites de los m intervalos de clase que se van a construir, con longitudes $C_1, C_2, ..., C_m$; de esta manera:

$$L_1 = L_0 + C_1$$

$$L_2 = L_1 + C_2$$

$$L_i = L_{i-1} + C_i$$

$$L_m = L_{m-1} + C_m$$

El primer límite inferior, L_0, debe escogerse de tal manera que sea un poco menor que el dato más pequeño; un criterio para definirlo es el siguiente:

Como los datos están registrados con una cifra decimal, se entiende que el instrumento de medición usado tiene una precisión de hasta las décimas de minuto. Puede decirse que los datos tienen (3) cifras significativas, lo cual indica que el registro "4.2 minutos" está representando cualquier valor real en el intervalo: (4.15 , 4.25), de esta manera puede definirse $L_0 = 4.15$.

Si se quiere que todos los intervalos de clase sean igual longitud, es decir $C_1 = C_2 = ... = C_m = C$, se deberá adoptar un valor C, que puede ser arbitrario o estimado con base en el rango de los datos. En este caso, una aproximación de C puede lograrse así:

$$C \cong \frac{\text{Rango}}{m}$$

Para el ejemplo 2.2 se construirán intervalos de diferente tamaño, por ser la situación más general.

Comenzando con $L_0 = 4.15$ podemos definir los otros límites como:

$L_1 = 7.15$, $L_2 = 11.15$, $L_3 = 13.15$, $L_4 = 16.15$, $L_5 = 18.15$, $L_6 = 21.15$, $L_7 = 27.15$, en este caso las longitudes de los 7 intervalos de clase son respectivamente 3, 4, 2, 3, 2, 3 y 6.

Para determinar la frecuencia asociada con cada intervalo, deben contarse los datos que pertenecen a cada uno; las definiciones de las frecuencias dadas anteriormente siguen vigentes para el caso de variables continuas, lo mismo que sus propiedades.

Se determina el punto medio de cada intervalo, que se denomina marca de clase y se representa por x'_i así:

$$x'_i = \frac{L_{i-1} + L_i}{2}$$

Este valor se constituye en el "representante" de los que pertenecen al intervalo correspondiente y más adelante jugará su papel.

A continuación se construye un cuadro de frecuencias para el ejemplo 2.2.

CUADRO 2.3

CUADRO DE FRECUENCIAS DE TIEMPOS DE ATENCION A PACIENTES EN EL "FILTRO" DEL SERVICIO DE URGENCIAS DE UN HOSPITAL
(En minutos)

Clase No. i	Intervalo de clase L_{i-1}-L_i	Marca de clase x_i	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia acumulada[1]	
					absoluta N_i	relativa F_i
1	4.15-7.15	5.65	2	0.04	2	0.04
2	7.15-11.15	9.15	5	0.10	7	0.14
3	11.15-13.15	12.15	6	0.12	13	0.26
4	13.15-16.15	14.65	15	0.30	28	0.56
5	16.15-18.15	17.15	9	0.18	37	0.74
6	18.15-21.15	19.65	8	0.16	45	0.90
7	21.15-27.15	24.15	5	0.10	50	1.00
			50	1.00		

OBSERVACIONES

1. Se puede apreciar en el cuadro 2.3. que el límite superior de un intervalo coincide con el límite inferior del siguiente, lo cual podría originar un problema de indefinición en caso de que un dato coincidiera con un límite, no se sabría donde clasificarlo. En el ejemplo no puede existir este problema puesto que todos los límites se han construido con una cifra decimal adicional a la que tienen los datos; cuando aquella posibilidad exista, se recomienda la convención: $(L_{i-1} , L_i]$ que significa que en cualquier intervalo de clase, el límite inferior no pertenece a él, pero sí, su límite superior.

2. Cuando los datos se agrupan en intervalos de clase, se produce pérdida de información, puesto que no se dispone de los datos en forma individual sino una caracterización más global, por ejemplo cuando se dice que en el intervalo 4.15 - 7.15 hay 2 datos, con ello no se sabe que valor tienen los dos datos, por tal razón cuando se reduce el número de intervalos se está globalizando más los datos y por tanto perdiendo más información. Por otro lado si se construyen demasiados intervalos se desvirtúa el objetivo de la estadística descriptiva, puesto que su manipulación se hace compleja y su presentación poco comprensible. Por tanto se recomienda que, en caso de que no exista una razón especial, se tome un número de intervalos mayor que cinco (5) y menor que veinte (20).

3. No deben existir intervalos de clase que no contengan datos. Con la distribución de frecuencias de la muestra se pretende explorar la distribución de la población; si existen clases sin datos se distorsiona esta idea. Cuando esto ocurra deberán reagruparse los datos.

4. Cuando sea posible debe procurarse que todos los intervalos sean de igual longitud, lo cual en ocasiones simplifica algunos cálculos y sobre todo facilita la interpretación, puesto que comparando directamente las frecuencias, se está comparando la densidad (concentración) en cada intervalo.

En algunas veces no es posible construir intervalos de igual longitud, por ejemplo, cuando la variable "salario" toma un rango amplio de valores, para bajos salarios, clases de $100.000 de longitud pueden considerarse, por ser esta diferencia importante, pero para altos salarios esta

longitud resulta pequeña. En estas situaciones la longitud de los intervalos crece con los valores de la variable, incluso a veces los intervalos extremos pueden ser abiertos ("los que ganan menos de $500.000" o los que ganan $1´000.000 o más).

Cuando los intervalos de clase son de diferente tamaño como en el ejemplo presentado, se dificulta conocer donde hay mayor concentración de los datos, esta situación se soluciona calculando la densidad de frecuencia relativa de cada intervalo, que consiste en expresar el porcentaje (o fracción) promedia de datos que hay por cada unidad de intervalo de clase.

Así por ejemplo el intervalo 13.15 - 16.15 contiene el 30% de los datos. Como el intervalo tiene una longitud de 3 minutos, se puede decir que dicho intervalo tiene una densidad promedio de 10% por cada minuto, que es el resultado de plantear: "si el 30% de los datos están en una longitud de 3 minutos, en un minuto que porcentaje habrá?

De esta manera si se asume que los datos en cada intervalo están uniformemente distribuidos, se puede definir la densidad f^*_i en el *i-ésimo* intervalo, como:

$$f_i^* = \frac{f_i}{C_i}$$

Si se expresa la densidad como una función para cualquier número real x, se obtiene la llamada **función empírica de densidad**, que para el ejemplo 2.2 estará dada por:

$$f^*(x) = \begin{cases} 0 & \text{si } x < 4.15 \text{ ó } x > 27.15 \\ 0{,}04/3 \equiv 1{,}33\%\ /\text{min} & \text{si } 4.15 < x \leq 7.15 \\ 0{,}10/4 \equiv 2{,}5\%/\text{min} & \text{si } 7.15 < x \leq 11.15 \\ 0{,}12/2 \equiv 6\%/\text{min} & \text{si } 11.15 < x \leq 13.15 \\ 0{,}30/3 \equiv 10\%/\text{min} & \text{si } 13.15 < x \leq 16.15 \\ 0{,}18/2 \equiv 9\%/\text{min} & 16.15 < x \leq 18.15 \\ 5.33\%\ /\text{min} & \text{si } 18.15 < x \leq 21.15 \\ 1.66\%\ /\text{min} & 21.15 < x \leq 27.15 \end{cases}$$

La palabra "empírica" es para resaltar que proviene de una muestra, pero pretende indicar el comportamiento de la variable en la población (función de densidad de probabilidad).

La expresión general para la **función empírica de densidad**, está dada por:^

$$f^*(x) = \begin{cases} 0 & x \leq L_0 \ \ x > L_m \\ \dfrac{f_i}{C_i} & L_{i-1} < x \leq L_i, \quad i = 1, 2, \ldots, m \end{cases}$$

Como puede apreciarse en la función empírica de densidad del ejemplo el intervalo

13.15 - 16.15 tiene la mayor concentración de datos (10 % /min).

2.2.1 Función empírica de densidad, $f^*(x)$.

Este gráfico es conocido con el nombre de histograma y consiste en una serie de rectángulos, cuya base son los intervalos de clase y su altura la densidad correspondiente.

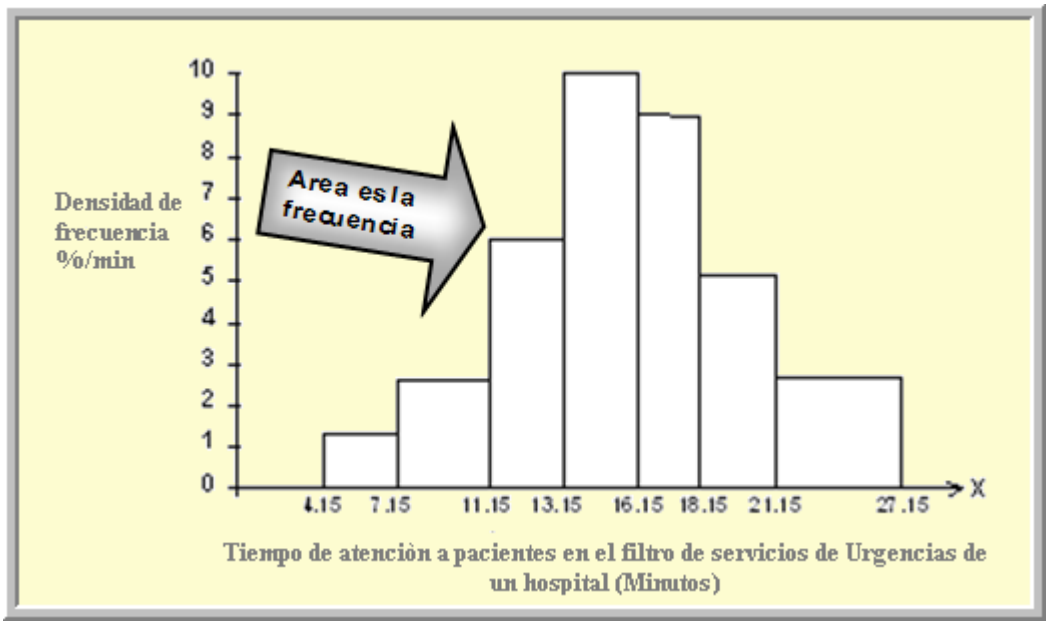


Fig. 2.3. *Histograma: gráfico de la función empírica de densidad.*

Al observar la figura 2.3 se puede apreciar que el área de uno de los rectángulos, por ejemplo el *i-ésimo* es:

$$A_i = base\ x\ altura$$

$$= C_i\ x\ f^*_i$$

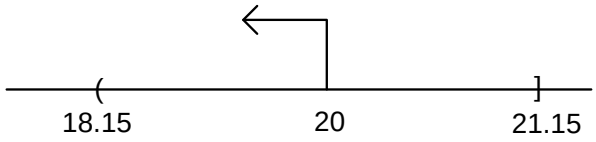
como $f_i^* = \frac{f_i}{C_i}$, entonces :

$$A_i = C_i \text{ x} \frac{f_i}{C_i} = f_i$$

Lo cual significa que el área de cada rectángulo es equivalente con su frecuencia relativa; de esta manera si un rectángulo tiene el doble de área que otro significa que contiene el doble de datos.

La suma de todas las áreas debe dar 100% ó 1.00.

La función empírica de densidad puede usarse para calcular en forma aproximada el porcentaje de datos que hay en un intervalo cualquiera. Si en el ejemplo 2.2 se deseara estimar el porcentaje total de consultas que duran 20 minutos o menos, se procede de la siguiente manera:



El porcentaje de datos menores o iguales que 20 puede calcularse al sumar el porcentaje de datos menores ó iguales a 18.15 (74%) más el porcentaje de datos que hay entre 18.15 y 20, el cual puede obtenerse mediante el siguiente razonamiento: "si en el intervalo 18.15 - 21.15 se tiene una densidad de 5.33 %/min entonces que porcentaje de los datos habrá en una longitud de (20 - 18.15) minutos?

$$\frac{5.33\%}{\text{min}}(20 - 18.15)\text{min} = 9.86\%$$

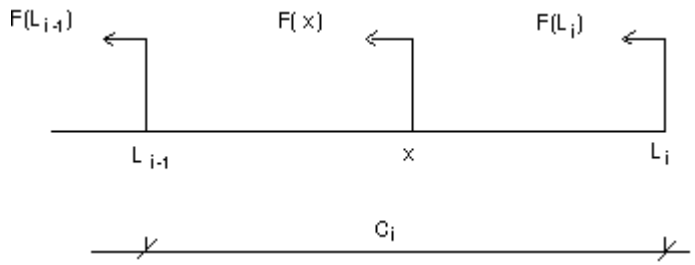
Así pues que el porcentaje de datos que son menores o iguales que 20 es:

$$F(20) = F(18.15) + 9.86\%$$

$$= 74\% + 9.86\% = 83.86\%$$

Con el mismo procedimiento se puede construir en forma general, para cualquier *x,* el porcentaje (o fracción) de datos que son menores o iguales que *x,* que se denota por *F(x)* y se conoce como **función empírica de distribución acumulativa**.

Supóngase que *x* pertenece al intervalo $(L_{i-1}, L_i]$ el cual tiene una longitud C_i y una frecuencia relativa f_i, e interesa conocer la frecuencia relativa acumulada hasta *x.*



En virtud del supuesto sobre la homogeneidad en la distribución de los datos en cada intervalo, se puede plantear la siguiente regla de tres: "si en C_i unidades hay una frecuencia f_i, en $(x - L_{i-1})$ unidades, qué frecuencia habrá ?", la respuesta es:

$$\frac{f_i}{C_i}(x - L_{i-1})$$

Por lo tanto:

$$F(x) = F(L_{i-1}) + \frac{f_i}{C_i}(x - L_{i-1})$$

Con esto se puede plantear la función empírica de distribución acumulativa como:

$$F(x) = \begin{cases} 0 & \text{para } x \le L_0 \\ F(L_{i-1}) + \dfrac{f_i}{C_i}(x - L_{i-1}) & \text{para } L_{i-1} < x \le L_i \\ 1 & \text{para } x > L_m \end{cases} \quad i = 1, 2, \ldots, m$$

Si se reemplaza $f_i^* = \dfrac{f_i}{C_i}$, se puede escribir:

$$F(x) = \begin{cases} 0 & \text{para } x \le L_0 \\ F(L_{i-1}) + f^*_i(x - L_{i-1}) & \text{para } L_{i-1} < x \le L_i \quad i = 1, 2 \ldots m \\ 1 & \text{para } x \ge L_m \end{cases}$$

La función de distribución acumulativa para el ejemplo 2.2, está dada por:

$$F(x) = \begin{cases} 0 & \text{si } x \le 4.15 \\ \frac{0.04}{3}(x - 4.15) & \text{Si } 4.15 < x \le 7.15 \\ 0.04 + \frac{0.10}{4}(x - 7.15) & \text{Si } 7.15 < x \le 11.15 \\ 0.14 + \frac{0.12}{2}(x - 11.15) & \text{Si } 11.15 < x \le 13.15 \\ 0.26 + \frac{0.30}{3}(x - 13.15) & \text{Si } 13.15 < x \le 16.15 \\ 0.56 + \frac{0.18}{2}(x - 16.15) & \text{Si } 16.15 < x \le 18.15 \\ 0.74 + \frac{0.16}{3}(x - 18.15) & \text{Si } 18.15 < x \le 21.15 \\ 0.90 + \frac{0.10}{6}(x - 21.15) & \text{Si } 21.15 < x \le 27.15 \\ 1 & \text{Si } x > 27.15 \end{cases}$$

Si se desea estimar el porcentaje de datos que son menores o iguales que 15 minutos, es decir:

$$F(15) = 0.26 + \frac{0.30}{3}(15 - 13.15)$$

$$= 0.26 + 0.185 = 0.445$$

O sea que el 44.5% de los pacientes son atendidos en 15 minutos o menos.

Si se desea estimar el porcentaje de datos que hay entre *"a"* y *"b"*, dígase *f(a,b)* se puede calcular como:

$$f(a,b) = F(b) - F(a)$$

Así por ejemplo, el porcentaje de datos que hay entre 15 minutos y 20 minutos puede estimarse como:

$$f(15;20) = F(20) - F(15) = 0.8386 - 0.445 = 0.3936$$

O sea que aproximadamente el 39.4% de los pacientes son servidos en el "filtro" en un tiempo entre 15 y 20 minutos.

2.2.2. Función empírica distribución acumulativa, *F(x)*.

De la función *F(x)* en el ejemplo 2.2, se observa que en cada intervalo, *F(x)*, representa un segmento de la recta, cuya pendiente es la densidad del intervalo respectivo. Esto da origen al siguiente gráfico con el nombre de **ojiva**.

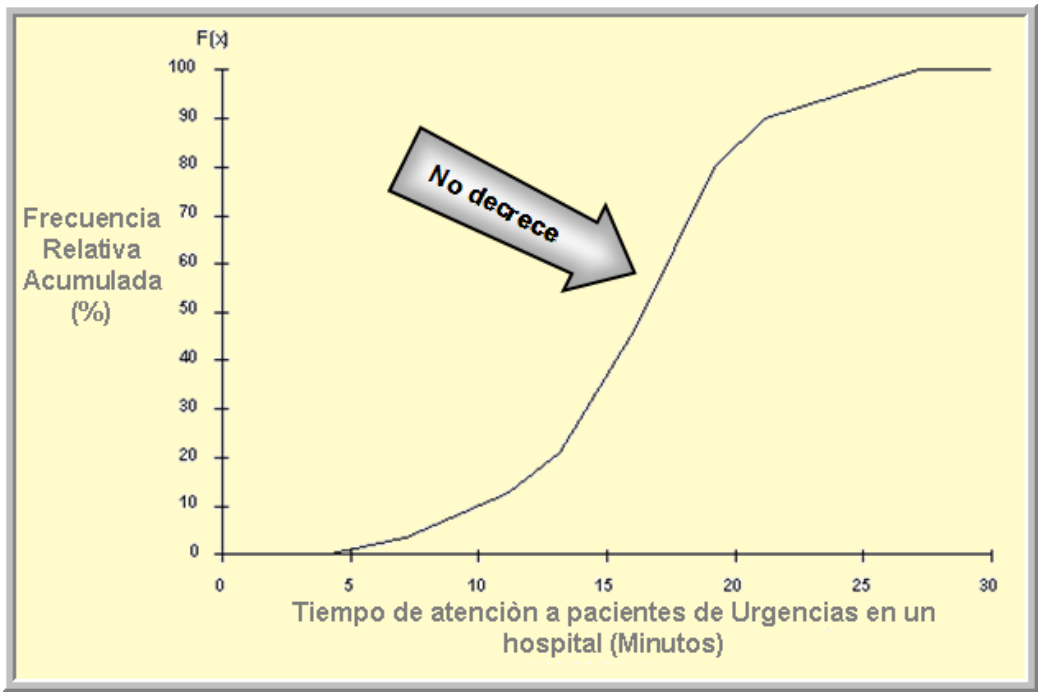


Fig. 2.4. *Ojiva: Función empírica de distribución acumulativa.*

Relación entre una función de densidad empírica y una función de densidad de probabilidad de las llamadas variables aleatorias Continuas.

Estas mismas ideas que se han desarrollado hasta ahora a partir de los datos de una muestra, tienen sus respectivos homólogos cuando se trabaja con todos los datos de la población estadística y las variables continuas con las que trabajamos recibirían el nombre de variables aleatorias, análogamente las funciones de densidad empíricas $f^*(x)$ y la Función de distribución acumulada F(x), reciben los nombres de función de densidad de probabilidad y Funcion de distribución acumulativa de probabilidad. Aquí intentaremos dar el paso de una manera natural de los conceptos de las muestras a los conceptos de las poblaciones, es decir, de las frecuencias relativas a la probabilidad y de las áreas de los rectángulos en el histograma a las áreas bajo curvas o funciones y en los cálculos pasaremos de las suma de áreas de rectángulos al calculo de intergrales. Ilustraremos este proceso con el siguiente ejemplo.

Ejemplo 2.2 B. (Del Histograma a función de densidad de Probabilidad)

En el sector de la industria metalmecánica, se toma una muestra al azar de 500 obreros y se determina la antigüedad en su trabajo.

Por razones de índole administrativo, se quiere representar los datos por medio de un histograma que considere los siguientes intervalos de clase: 0-2 años, 2-3 años, 3-5 años, 5-10 años, 10-20 años.

i	Intervalo (Años de Antigüedad)	Frecuencia Relativa % (f_i)
1	0-2	10%
2	2-3	5%
3	3-5	40%
4	5-10	40%
5	10-20	5%
	TOTAL	100%

Cuadro 2B1. Distribución de frecuencias de la Antigüedad en el trabajo.

Los intervalos del cuadro, incluyen el límite superior, pero no el inferior.

Observe que la frecuencia relativa la hemos denotado por f_i

Vamos a construir un histograma con los datos agrupados presentados en el cuadro..

Recordando las Bases para la construcción de un histograma.

Un histograma es una serie de rectángulos construidos cada uno de los cuales tiene como base el intervalo correspondiente y cuya área representa la frecuencia relativa f_i de su intervalo respectivo. De tal manera que un intervalo que contiene el doble de datos que otro, deberá estar representado por rectángulo que tiene el doble del área. (Ojo que se dice el doble de área y no de altura). Observe del cuadro de frecuencias de nuestro ejemplo, que el primer rectángulo, deberá tener el doble de área que el segundo. El Tercero deberá tener la misma área del cuarto y además debe tener 4 veces el área del primero, pues esa es la relación de las áreas.

Con estos criterios construyamos nuestro histograma.

Vamos a construir el primer rectángulo de un área arbitraria, pero las demás áreas deberán guardar proporcionalidad de acuerdo con las frecuencias relativas f_i .

Si vemos el gráfico de la Figura, se aprecia muy claramente la proporcionalidad de las áreas de acuerdo con la frecuencia relativa de cada intervalo. Observe por ejemplo que el primer rectángulo tiene el doble de área que el segundo, no obstante que tienen la misma altura. Note como los intervalos tercero y cuarto tienen rectángulos con la misma área, no obstante que las alturas son distintas. También el primero y el último tienen la misma área, pues en ambos hay el 5% de los datos.

Interpretación de la altura f_i^* de los rectángulos de un histograma.

Si el área representa la frecuencia relativa (% de datos), entonces como se puede interpretar la altura de un rectángulo? Qué significado tiene el valor de la altura de uno de los rectángulos del histograma?.

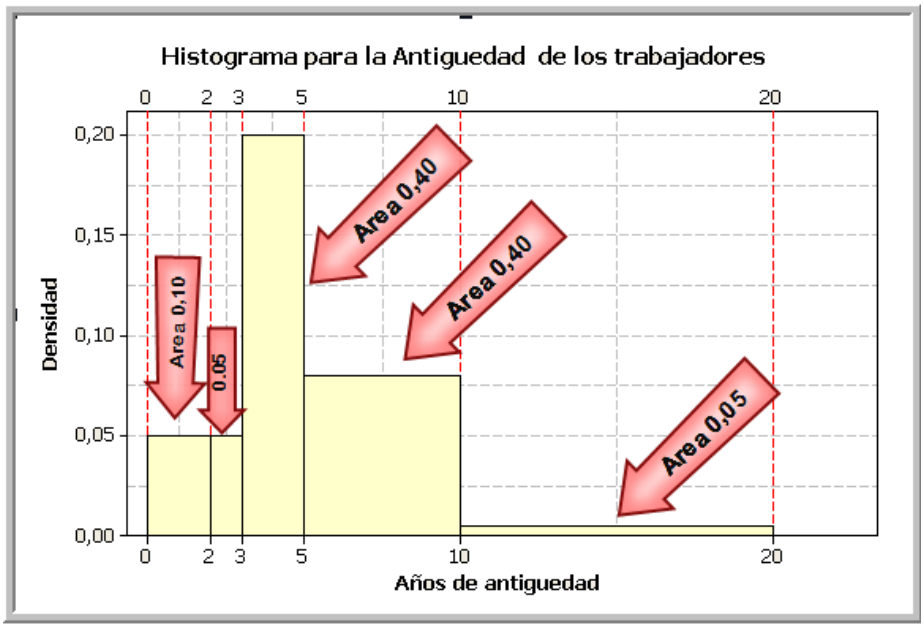


Figura 2.4B. Histograma para la variable "Antigüedad en el Trabajo"

Por lo pronto denotemos la altura del rectángulo i-esimo, por f_i^*, observe que le hemos colocado un (*) para diferenciarlo de f_i .

Llamemos C_i al ancho del intervalo i. De esta manera $C_1 = 2$, $C_2 = 1$, $C_3 = 2$, $C_4 = 5$, $C_5 = 10$

De la definición de histograma quedó establecido que las áreas representan las frecuencias relativas respectivas, es decir que si llamamos A_i al área correspondiente, entonces estamos diciendo que: $A_i = f_i$, pero como el área de un rectángulo es base por altura, entonces:

$A_i = f_i = base * altura = C_i * f_i^*$, de donde podemos calcular f_i^* , despejando obtenemos:

$f_i^* = \frac{f_i}{C_i}$. Observe que se divide la frecuencia relativa entre el número de unidades que tenga el intervalo correspondiente, entonces las unidades de f_i^* son (% de datos por cada unidad de la variable en dicho intervalo). Veamos por ejemplo para el primer intervalo: $f_1 = 10\%$ y $C_1 = 2$, así

que la altura del primer rectángulo es: $f_1^* = \frac{f_1}{C_1} = \frac{10\%}{2\ años} = 5\%/año$, que escrito en forma decimal es 0.05/año. (vea la Figura.2.4B).

Es intuitivamente claro, que si el primer intervalo tiene el 10% de los datos y estos datos están distribuidos en un intervalo que tiene una longitud de dos (2) unidades, pues en promedio hay 5% por cada unidad ($f_1^* = 5\%/año \equiv 0.05/año$)

El cuarto intervalo, (5; 10], por ejemplo, en sus 5 unidades (5 años) contiene 40% de los datos. Así que en promedio, hay 8% de los datos en cada unidad o lo que es lo mismo:

$$f_4^* = \frac{f_4}{C_4} = \frac{40\%}{5\ años} = 8\%/año \equiv 0,08/año$$

Es decir que las unidades del eje Y en el gráfico de la Figura.2.4B, es 1/unidad o %/unidad, por eso se le conoce como densidad de frecuencia (f_i^*).

i	Intervalo (Años de Antigüedad)	Frecuencia Relativa % (f_i)	Densidad de Frecuencia (f_i^*)
1	0-2	10%	5%/año
2	2-3	5%	5%/año
3	3-5	40%	20%/año
4	5-10	40%	8%/año
5	10-20	5%	0,5%/año
	TOTAL	100%	

Cuadro 1B2. Densidad de frecuencia para la antigüedad en el trabajo.

En general, si queremos estimar el porcentaje de datos que hay en cualquier intervalo de antigüedad, solo deberemos calcular su área asociada en el histograma. Veamos un ejemplo:

¿Cuál es el porcentaje de obreros que tienen antigüedad menor que 4 años?.

Este porcentaje corresponde al área sombreada en la figura:

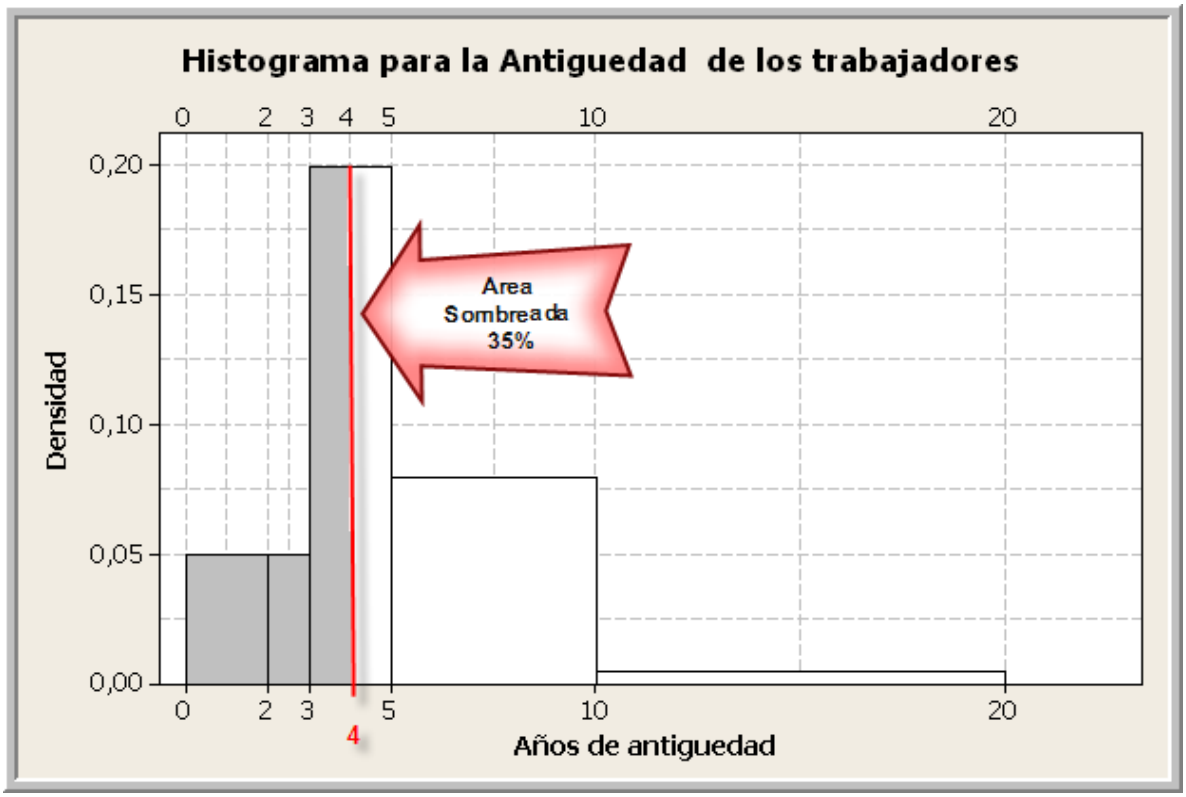


Figura 2.4C. Representación del porcentaje de trabajadores con antigüedad de 4 años o menos.

Observe que el área sombreada se calcula sumando por un lado las áreas de los primeros rectángulos (10%+5%) y por otro lado la parte del tercer rectángulo comprendida entre 3 y 4, que resulta ser la mitad de 40%, es decir 20%. Así que el porcentaje de trabajadores con antigüedad de 4 años o menos se estima en:

$$P(X \leq 4) = 10\% + 5\% + 20\% = 35\% \equiv 0,35$$

Haciendo cuentas usando el concepto de densidad de frecuencia, podríamos decir que como en el tercer intervalo su densidad es de 20%/año y en entre 3 y 4 años hay una unidad, entonces habrá el 20%.

Estimemos ahora el porcentaje de trabajadores con antigüedad entre 4 y 7,5 años.

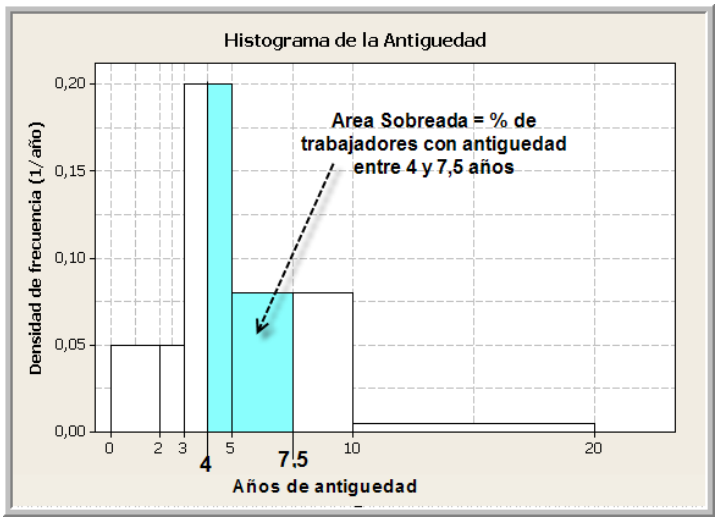


Figura2.4D. Representación en el Histograma del porcentaje de trabajadores con Antigüedad entre 4 y 7,5 años.

$$P(4 \le X \le 7{,}5) = f_3^* * (5-4) + f_4^* * (7{,}5-5) = 20\%/año * (1año) + 8\%/año * (2{,}5años) = 40\%$$

Rec

uerde que el eje Y (altura de los rectángulos) representan la densidad de frecuencia *f**

Observe que el área total del histograma siempre será 100%.

Si un valor x_0 se encuentra en el cuarto intervalo, es decir entre 5 y 10. Encuentre el porcentaje de trabajadores con antigüedad menor o igual que x_0.

De la Figura.2.4E, se puede apreciar al calcular el área acumulada hasta x_0, que:

$$P(X \le x_0) = 10\% + 5\% + 40\% + 8\%/año * (x_0 - 5) =$$

$$P(X \le x_0) = 55\% + 8\%/año * (x_0 - 5)$$

Aquí hemos obtenido una fórmula para calcular la frecuencia relativa acumulada hasta x_0, cuando este valor se encuentra entre 5 y 10 años de antigüedad.

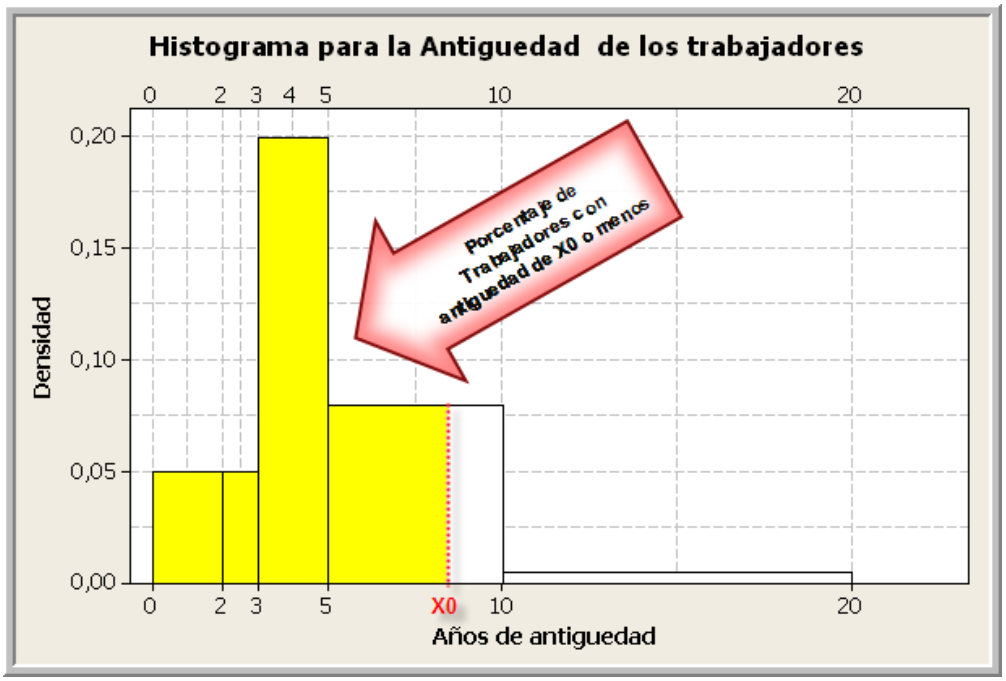


Figura 2.4E. Representación del porcentaje de Trabajadores con antigüedad de x_0 o menos

Así pues si x_0=8 años, entonces: $P(X \le 8) = 55\% + 8\%/año * (8-5) años = 79\%$.

Si cada vez cambiamos el intervalo en el cual se encuentra x, podemos obtener la siguiente función F(x), para calcular $P(X \le x)$.

$$F(x) = P(X \le x) = \begin{cases} 0 & x \le 0 \\ 0,05 * x & 0 < x \le 2 \\ 0,10 + 0,05 * (x-2) & 2 < x \le 3 \\ 0,15 + 0,20 * (x-3) & 3 < x \le 5 \\ 0,55 + 0,08 * (x-5) & 5 < x \le 10 \\ 0,95 + 0,005 * (x-10) & 10 < x \le 20 \\ 1 & x > 20 \end{cases}$$

Función de Distribución de Frecuencia Relativa Acumulada.

Examine la expresión obtenida para F(x)= $P(X \le x)$ y asegúrese de saber construirla.

Usando dicha expresión podemos estimar por ejemplo el porcentaje F(4), es decir el porcentaje de trabajadores con 4 años de antigüedad o menos: Observe que x=4, se encuentra en el intervalo $3 < x \leq 5$, por lo tanto:

$$F(4) = P(X \leq 4) = 0,15 + 0,20 * (4 - 3) = 0,35 \equiv 35\%$$

Ahora imaginemos que disponemos de un número muy grande de datos de tal manera que sea posible construir muchos intervalos de pequeña anchura y a tal punto que el conjunto de rectángulos del histograma se convierte en una curva suave $f^*(x)$ como se muestra en la Figura . El área sombreada ilustra a F(x)= $P(X \leq x)$.

Note que si ahora conociéramos la expresión para $f^*(x)$, el área sombreada podría calcularse como:

$F(x) = P(X \leq x) = \int_{-\infty}^{x} f^*(x).dx$ es decir, que el área ahora podría calcularse como la integral bajo la curva.

A esta función suave $f^*(x)$ que se supone ahora describe la población completa y no una muestra le llamaríamos *función de densidad de probabilidad de la variable aleatoria antigüedad.*

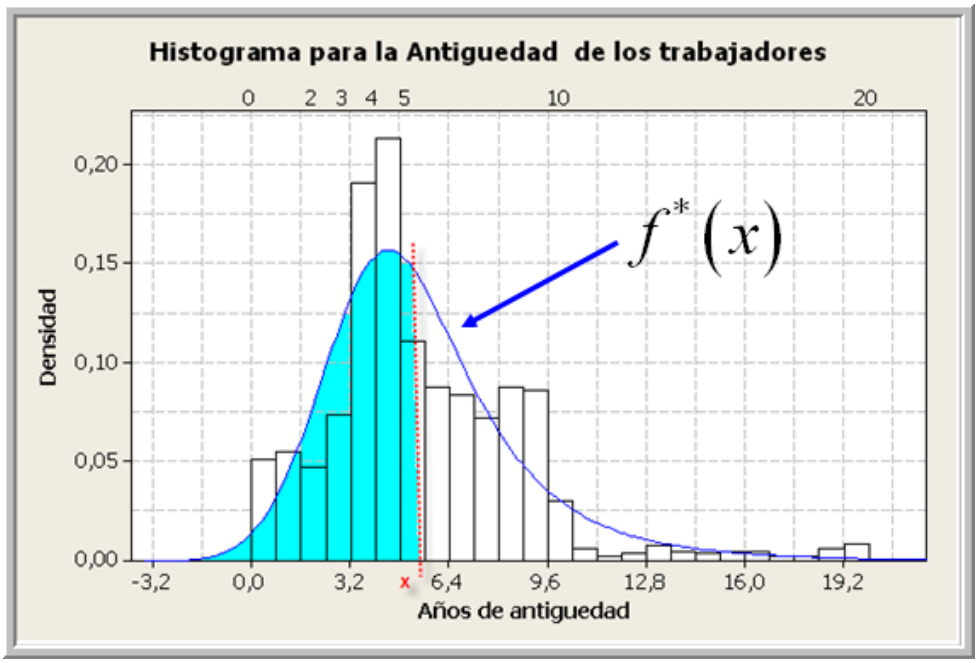


Figura 2.4F. Idealización de una función de densidad de probabilidad

Ahora estamos preparados para la definición de variable aleatoria continua.

Variable aleatoria continua. Definición.

Se dice que X es una variable Aleatoria Continua si existe una función f(x), llamada función densidad de probabilidad (fdp) de X, que satisface las siguientes condiciones:

a) $f(x) \geq 0 \quad \forall x \in \Re$

Es razonable que no tome valores negativos, siendo una función de densidad de probabilidad.

b) $\int_{-\infty}^{+\infty} f(x).dx = 1$

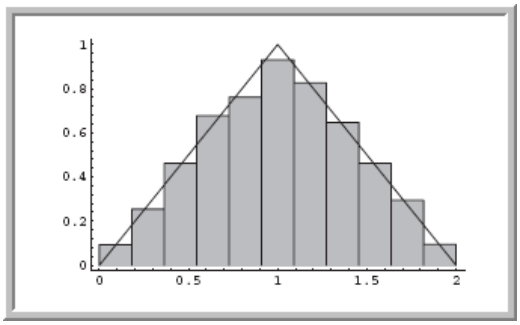
Ya hemos dicho antes que el área del histograma y ahora el área bajo la función de densidad, debe ser 100%.

c) Para cualquier a, b se tiene que $P(a \leq X \leq b) = \int_{a}^{b} f(x).dx$

El área atrapada entre los valores a y b es justamente el porcentaje de datos de la población que cumple con esas especificaciones. Mirado como la experiencia aleatoria de sacar al azar un valor de X, esta área puede interpretarse como probabilidad.

Ejemplo 2.2C.

El Histograma de una cierta característica continua X, es el que muestra sombreado en la figura.



Se pretende ajustar una función densidad y suena razonable la que aparece ajustada formando un triangulo equilátero. Encuentre la definición de dicha función de densidad de probabilidad estimada, f(x).

En primer lugar se observa que el rango de valores que puede tomar la variable aleatoria X son los puntos en el intervalo que va de cero (0) a dos(2). Es decir que:

$$\Omega_X = \{x \in \Re / 0 \prec x \leq 2\}$$

Rango o Recorrido de la variable aleatoria X. algunas veces se denota por $\Re_X$

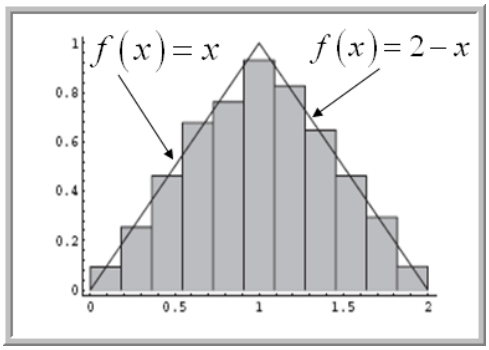
Cual deberá ser la ecuación que defina las dos rectas que conforman el triangulo equilátero y que definen la función de densidad de probabilidad estimada?.

Pues como el área debe ser igual a la unidad, esto significa que la altura h del triangulo, debe ser tal que el área valga 1.

$$Area = 1 = \frac{base * altura}{2} = \frac{2 * h}{2} = 1$$

De donde se deduce que la altura h=1. Por lo tanto la ecuación de la recta de pendiente positiva es f(x)=x. la ecuación de la recta con pendiente negativa será: f(x)=2 –x, así pues:

$$f(x)=\begin{cases} x & 0<x\le 1 \\ 2-x & 1<x\le 2 \end{cases}$$



Si se produce una realización de la variable aleatoria X, estime la probabilidad de que el valor resulte entre 0,5 y 1,5?

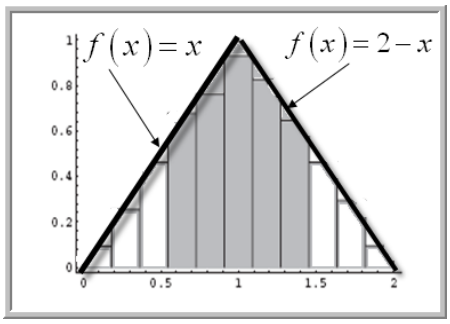
$$P(0,5\le X\le 1,5)=\int_{0,5}^{1,5} f(x).dx$$

$$P(0,5\le X\le 1,5)=\int_{0,5}^{1,0} x.dx+\int_{1,0}^{1,5}(2-x).dx=$$

$$P(0,5\le X\le 1,5)=\int_{0,5}^{1,0} x.dx+\int_{1,0}^{1,5}(2-x).dx=$$

$$P(0,5\le X\le 1,5)=\frac{x^2}{2}\Big|_{0,5}^{1,0}+\left(2x-\frac{x^2}{2}\right)\Big|_{1,0}^{1,5}=$$

$$P(0,5\le X\le 1,5)=\frac{3}{4}$$



Observe que el área, en este caso, se hubiera podido calcular como el área de dos trapecios, con base mayor la altura del triangulo.

Ejemplo 2.2D

El tiempo, en horas, que tarda un autobús urbano en completar su recorrido se puede representar mediante una variable aleatoria X con la siguiente función de densidad:

$$f(x)=\begin{cases} \text{kx} & ;\quad 0\le \text{x}\le 1 \\ 0 & ;\quad \text{resto} \end{cases}$$

Obtener el valor de k para que f(x) sea una función de densidad.

De acuerdo a las propiedades de una función de densidad para variables aleatorias continuas se tiene que: $f(x) \geq 0$ y además $\int_{-\infty}^{\infty} f(x)dx = 1$

Es decir que $\int_0^1 kxdx = 1$, por lo tanto:

$$\Rightarrow \int_0^1 kxdx = k\int_0^1 xdx = k\left[\frac{1}{2}x^2\Big|_0^1\right] = \frac{k}{2}\left[(1)^2 - (0)^2\right] = \frac{k}{2}(1) = \frac{k}{2}$$

Ahora al igualar y despejar k se obtiene que: $\frac{k}{2} = 1 \Rightarrow k = 2$

Por lo tanto:

$$f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{en otra parte} \end{cases}$$

Obtener la función de distribución (Acumulada).

$$F(X) = P(X \leq x) = \int_0^x f(t)dt$$

$$\text{F(x)=P}\left(\text{X} \leq \text{x}\right) = \int_0^x 2tdt = 2\int_0^x tdt = 2\left[\frac{1}{2}t^2\right]\Big|_0^x = 2\left[\frac{1}{2}\left(x^2 - (0)^2\right)\right] = \frac{2}{2}\left(x^2\right) = x^2$$

Función de Distribución Acumulativa de Probabilidad

$$\text{F(x)=P}\left(\text{X} \leq \text{x}\right) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

¿Cuál es la probabilidad de que el autobús efectúe su recorrido como mucho en 3/4 de hora? ¿Y la probabilidad de que tarde más de 3/4 de hora?

La probabilidad de que el autobús efectúe su recorrido como mucho en 3/4 de hora se obtiene así:

$$F(3/4) = P\left(X \le 3/4\right) = \left(\frac{3}{4}\right)^2 = \frac{9}{16} = 0.5625$$

Lo cual significa que aproximadamente el 56% de las veces el autobús se tarda ¾ de hora o menos.

La probabilidad de que tarde más de 3/4 de hora es: $1 - F(X = 3/4) = 1 - 0.5625 = 0.4375$

Calcular la probabilidad de que el autobús tarde entre 20 minutos (1/3 de hora) y 1 hora en completar su recorrido.

Observe que $P\left(a \le X \le b\right) = P\left(X \le b\right) - P\left(X \le a\right) = F(b) - F(a)$

Por lo tanto: $P\left(\frac{1}{3} \le X \le 1\right) = P\left(X \le 1\right) - P\left(X \le \frac{1}{3}\right) = F(1) - F(\frac{1}{3})$

$$F(1/3) = P\left(X \le 1/3\right) = \left(\frac{1}{3}\right)^2 = \frac{1}{9} = 0.1111 \qquad F(1) = P\left(X \le 1\right) = \left(1\right)^2 = 1$$

Al hacer la diferencia se obtiene la probabilidad deseada.

$$F(1) - F(1/3) = 1 - 0.1111 = 0.8888$$

Por lo tanto la probabilidad de que el autobús tarde entre 20 minutos (1/3 de hora) y 1 hora en completar su recorrido es de 0.8888. Es decir que se espera que aproximadamente el 88,9% de las veces el autobús tarde un tiempo comprendido entre 20 minutos y una hora.

Ejemplo 2.2E

La duración de la tramitación de un expediente administrativo de licencia de obras es una variable aleatoria con distribución Exponencial , es decir con función de densidad de la forma

$f(x) = áe^{-áx}$; $x > 0$. De datos de experiencias anteriores se ha estimado que á=1/3.

Es decir que $f(x) = \frac{1}{3}e^{-\frac{1}{3}x}$; $x > 0$

Cierto constructor trabaja con avales bancarios para cada una de sus obras, de forma que los intereses que debe pagar empiezan a resultarle muy gravosos cuando las licencias sufren retrasos superiores a 4 meses. En estos momentos, el constructor tiene en proyecto un total de 12 obras. Calcule:

a) *La probabilidad de que una obra específica le resulte gravosa.*

En realidad lo que se pide es la probabilidad de que el tiempo de tramitación de una obra sea superior a 4 meses. P(X>4).

$$P(X > 4) = \int_4^{+\infty} f(x)dx = \int_4^{+\infty} \tfrac{1}{3}e^{-\frac{1}{3}x}dx$$

$$P(X > 4) = \int_4^{\infty} \frac{1}{3}e^{-\frac{1}{3}x}dx = \frac{1}{3}\int_4^{\infty} e^{-\frac{1}{3}x}dx = -e^{-\frac{1}{3}x}\Big|_4^{\infty} = -e^{-\frac{1}{3}(\infty)} + e^{-\frac{1}{3}(4)} = (0) + e^{-\frac{4}{3}} = 0.2635$$

Es decir que un poco más de la cuarta parte de las veces que se hace un trámite de licencia, ésta tarda más de 4 meses y resulta gravosa para el constructor

Ejemplo 2.2F

El porcentaje de alcohol (100X) en cierto compuesto se puede considerar como una variable aleatoria donde X, con la siguiente función de densidad de probabilidad:

$f(x) = 20x^3(1-x)$; $0 \leq x \leq 1$.

a) *Construya la Función F(x) de Distribución Acumulativa de Probabilidad.*

$$F(x) = P\left(X \le x\right) = \int_0^x f(x)dx$$

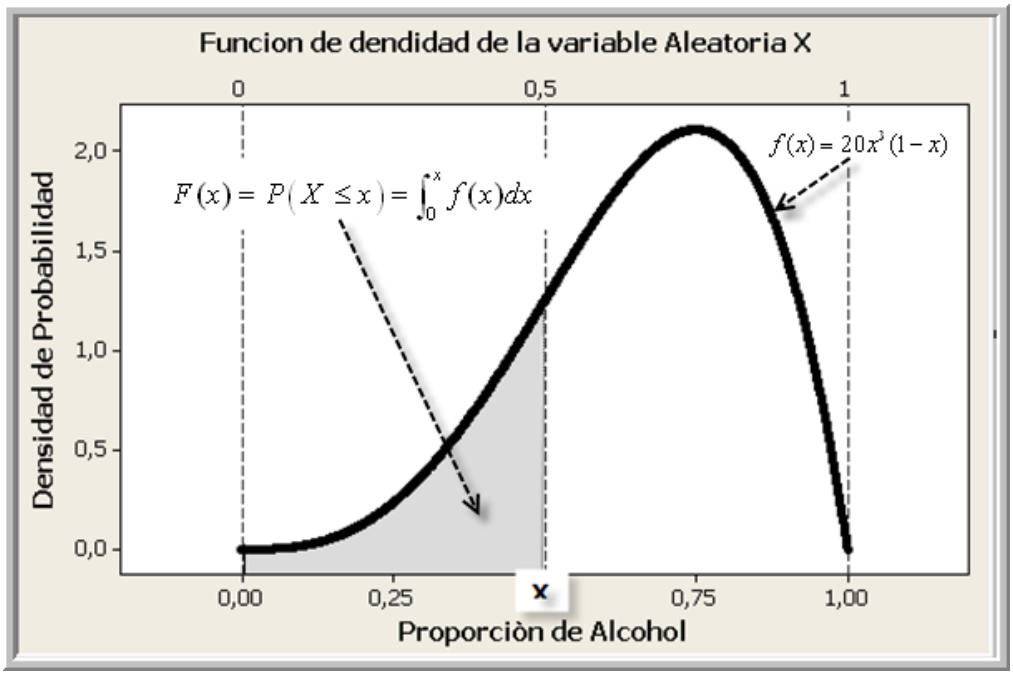


Figura 2.4G. Relación entre la Función de densidad de Probabilidad y la función de Distribución Acumulativa de Probabilidad F(x).

$$F(x) = \begin{cases} 0 & x < 0 \\ \int_0^x 20x^3(1-x)dx = 20\left(\frac{1}{4}x^4 - \frac{1}{5}x^5\right) & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 20\left(\frac{1}{4}x^4 - \frac{1}{5}x^5\right) & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

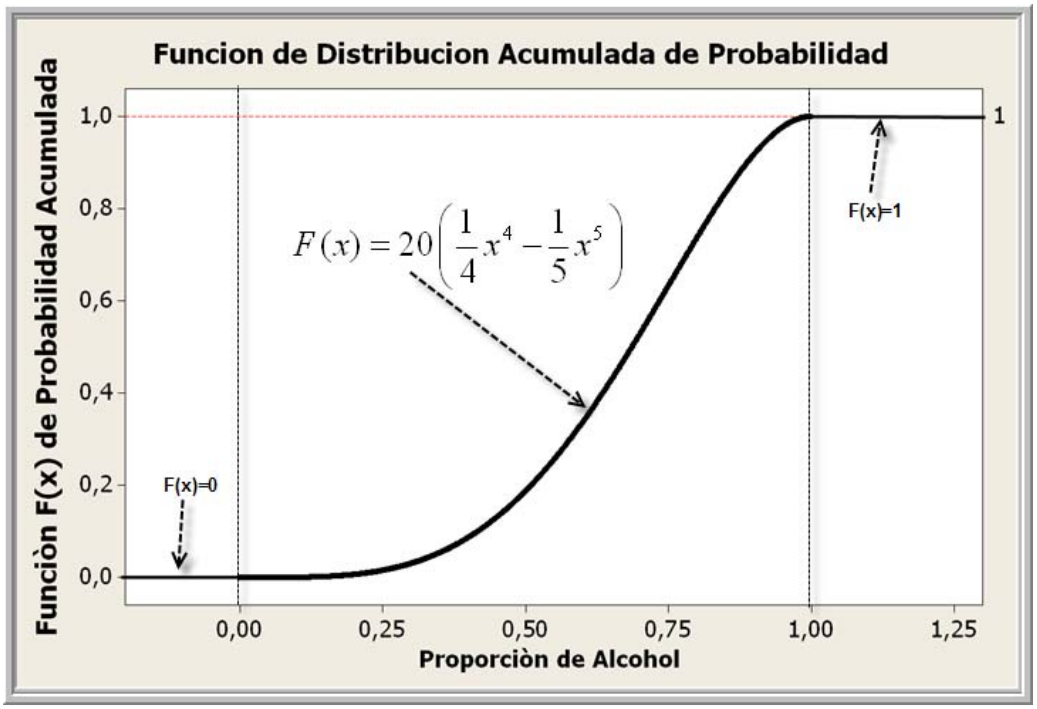


Figura2.4H. Función de Distribución Acumulativa de Probabilidad.

Observe que la Función de Distribución Acumulativa de Probabilidad, es no decreciente, lo cual es razonable, siendo que $F(x) = P(X \le x) = \int_0^x f(x)dx$ puesto que entre mayor sea x, mayor será el área bajo la función de densidad, o por lo menos no disminuye. Además note que está definida para todos los números reales.

b) *Calcule la probabilidad de que el compuesto contenga las dos terceras partes o menos de alcohol.*

$$P(X \le \tfrac{2}{3}) = F(\tfrac{2}{3}) = 20\left(\frac{1}{4}\left(\tfrac{2}{3}\right)^4 - \frac{1}{5}\left(\tfrac{2}{3}\right)^5\right) = 0{,}469$$

c) *Calcule el contenido mediano de alchol, es decir la mediana de la variable aleatoria X.*

Ya sabemos que la mediana es aquel valor $\tilde{x}$, tal que $P(X \leq \tilde{x}) = 50\%$, es decir aquel valor para el cual $F(\tilde{x}) = 0,50$, con lo cual:

$$20\left(\frac{1}{4}\tilde{x}^4 - \frac{1}{5}\tilde{x}^5\right) = 0,50$$

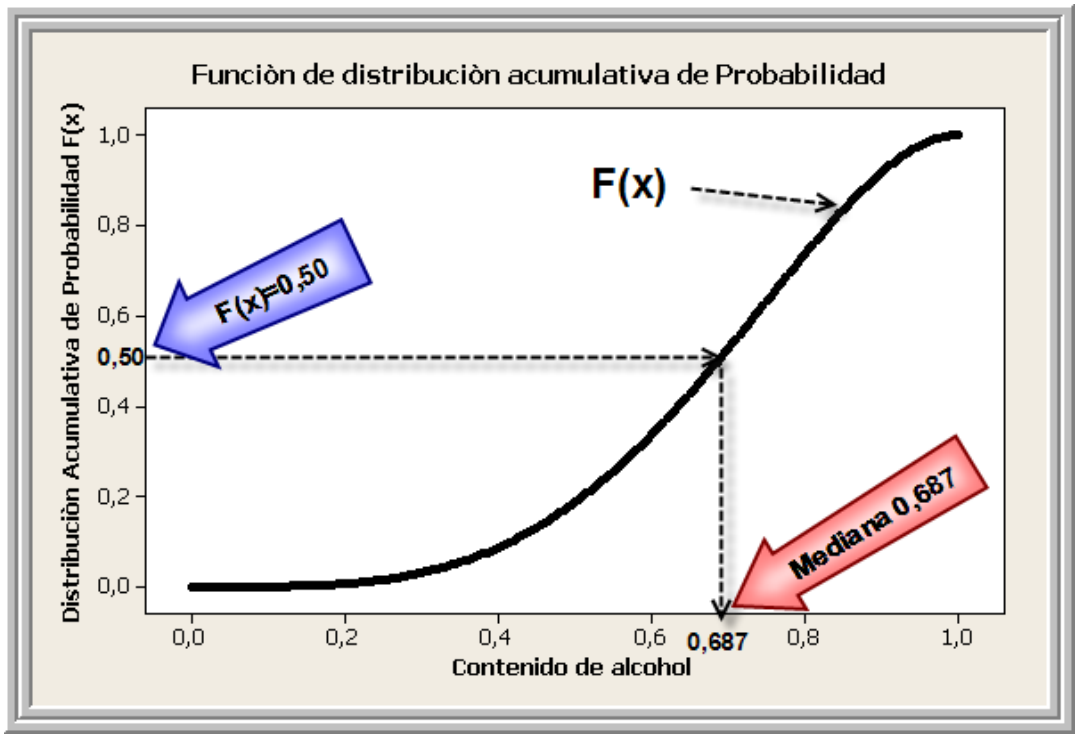


Figura 2.4I. Interpretación de la mediana de una variable aleatoria

Lo cual significa que la mediana del contenido de alcohol es 0,687, es decir que la mitad de las veces el compuesto resulta con 68,7% de alcohol o menos.

d) Supóngase que el precio de venta del compuesto anterior depende del contenido de alcohol. Específicamente si 1/3 ≤ X ≤2/3, el compuesto se vende a 50 dólares/galón, de otro modo se vende a 30 dólares /galón. Si el costo por galón del compuesto es 20 dólares /galón, entonces a la larga, cuanta es en promedio la utilidad por galón?

Definamos una nueva variable aleatoria que represente la Utilidad U, por galón.

$$U = \begin{cases} \$\ 30 & Si\ \frac{1}{3} \le X \le \frac{2}{3} \\ \$\ 10 & En\ Otro\ caso \end{cases}$$

Cuál es la distribución de probabilidad de la variable aleatoria Utilidad?

$$P(U = \$\ 30) = P(\tfrac{1}{3} \le X \le \tfrac{2}{3}) = F(\tfrac{2}{3}) - F(\tfrac{1}{3}) =$$

$$P(U = \$\ 30) = 20\left[\frac{1}{4}\left(\frac{2}{3}\right)^4 - \frac{1}{5}\left(\frac{2}{3}\right)^5 - \frac{1}{4}\left(\frac{1}{3}\right)^4 + \frac{1}{5}\left(\frac{1}{3}\right)^5\right] = 0,4156$$

Por lo tanto la $P(U = \$\ 10)$ será su complemento.

$P(U = \$\ 10) = 1 - 0,4156 = 0,5844$. En síntesis la distribución de probabilidad de la variable aleatoria Utilidad, U, es:

Utilidad U	Probabilidad
$ 30	0,4156
$ 10	0,5844
TOTAL	1,0000

Cuadro 2. Distribución de la variable aleatoria "Utilidad"

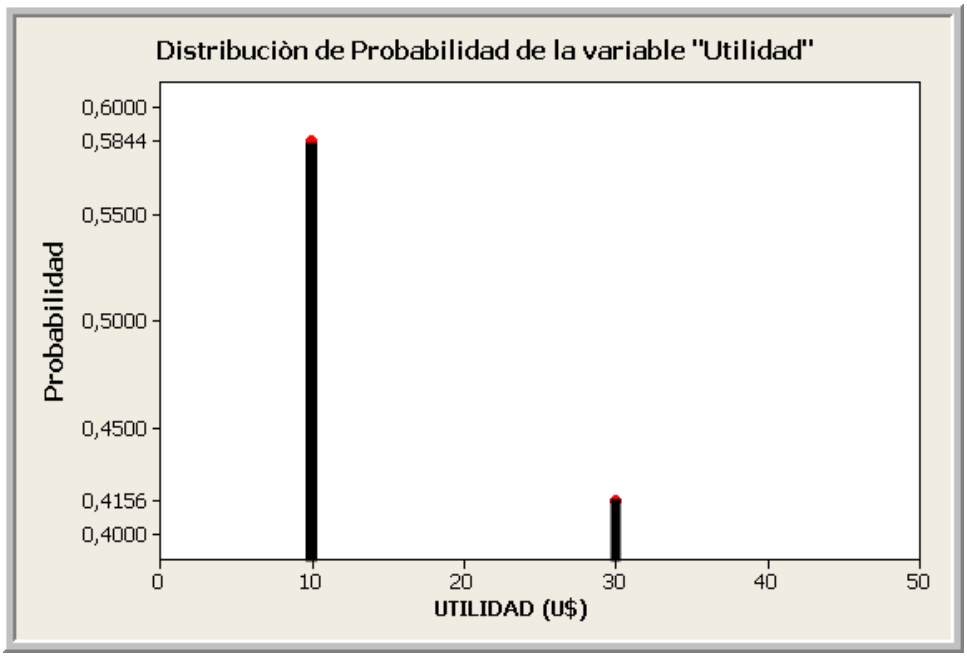


Figura 2.4J Distribución de Probabilidad de la variable Utilidad (U)

Camello 1 (trabajo para los estudiantes)

La Duración en horas de cierto dispositivo electrónico es una variable muy importante para una industria de productos electrónicos. Por esta razón se llevan muchos registros sobre la duración de dispositivos en experimentación.

6,322	121,631	137,359	151,732	191,043	241,303	249,998	346,725	913,010	1612,83	314
95,383	116,001	134,003	164,843	180,256	211,515	265,146	347,854	852,551	1039,55	490
105,782	117,280	143,637	153,579	181,115	231,737	283,384	473,724	725,766	1140,30	2317
98,223	116,044	137,679	160,832	172,031	238,263	254,300	398,529	822,989	1280,59	4825
110,449	116,277	131,399	152,870	184,158	203,605	305,500	353,028	734,850	1657,00	4364
94,534	121,007	130,067	149,850	172,533	216,436	287,826	348,263	883,315	1369,29	3890
101,697	122,751	134,079	160,227	185,842	205,348	294,900	429,046	681,405	1704,75	478
0,220	121,137	137,037	162,362	178,469	235,593	258,583	396,044	970,940	1589,00	2029
1,903	118,368	131,169	162,447	170,339	212,484	294,653	421,063	650,104	1032,07	253
98,862	119,709	135,323	150,424	184,567	239,609	251,582	351,439	700,056	1339,27	3081
109,646	120,951	137,438	163,282	168,396	229,204	275,650	425,423	792,997	1152,71	2704
110,280	115,866	141,597	145,680	198,797	240,179	311,564	339,350	505,555	1926,64	3243
6,604	115,436	133,965	152,683	180,632	224,139	250,205	328,088	659,109	1545,44	3301

Figura 2.4K. Registro de datos sobre la duración en horas de un dispositivo electrónico.

Con base en este gran conjunto de datos, se construyó un histograma que nos permite tener una estimación empírica de la función densidad, la cual está representada por las alturas de los rectángulos. Con base en la densidad empírica se ajustó el modelo que muestra la Figura que resultó ser $f(x)=\frac{100}{x^2} \quad x>100$.

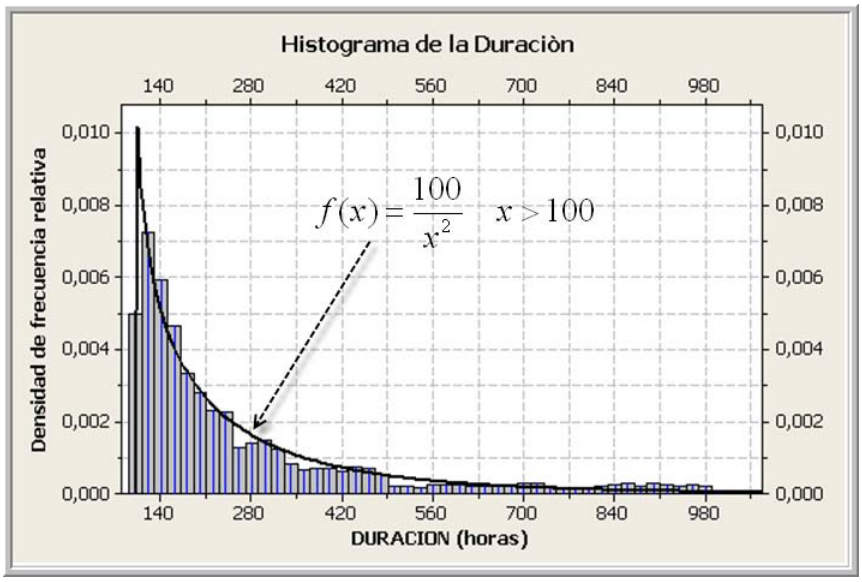


Figura 2.4 L. Ilustración del ajuste de un modelo para función de densidad de probabilidad

Con base en dicha función de densidad ajustada: a) Verifique que f(x) es una verdadera función de densidad b) Construya la Función de distribución acumulada de probabilidad para la duración. c) Estime la probabilidad de que un dispositivo dure menos de 200 horas. d) Estime la probabilidad de que un dispositivo dure más de 200 horas, si se sabe que todavía funciona después de 150 horas. e) De acuerdo con los resultados anteriores, decida si es razonable pensar que los dos eventos son independientes. f) Si se instalan 3 de estos dispositivos en un sistema y la duración de cada dispositivo es independiente de las de los otros, estime la probabilidad de que al menos uno de ellos dure más de 150 horas. g) Cuál es el número máximo "n" de dispositivos que deberán ponerse en un conjunto de modo que haya una probabilidad 0,50 de que después de 150 horas todos estén funcionando

Ejemplo 2. 2G.

Si un instrumento electrónico tiene una duración X (en unidades de 1000 horas) que se considera una variable aleatoria continua con función de densidad de probabilidad.

$f(x) = e^{-x}$ Para valores positivos de x.

El costo del artículo es $2, sin embargo el fabricante vende el artículo en $5, con la condición de que devuelve todo el dinero si el instrumento dura 900 horas o menos , es decir si X≤ 0,900.

a) *¿Cuál es la Función de distribución acumulada de probabilidad, F(x), para la variable aleatoria duración?*

$$F(x) = P\left(X \le x\right) = \int_0^x f(x)dx$$

$$F(x) = P\left(X \le x\right) = \begin{cases} 0 & Si\ x \le 0 \\ \int_{-\infty}^{x} e^{-x}dx = \underbrace{\int_{-\infty}^{0} e^{-x}dx}_{0} + \int_0^x e^{-x}dx & x > 0 \end{cases}$$

$$F(x) = P\left(X \le x\right) = \begin{cases} 0 & Si\ x \le 0 \\ 1 - e^{-x} & x > 0 \end{cases}$$

Función de Distribución Acumulativa de Probabilidad para la variable aleatoria duración, X.

b) *Calcule la probabilidad de que el fabricante deba devolver el dinero de la venta de un instrumento.*

En realidad la probabilidad pedida es P(X≤0,900), es decir F(0,900)

Probabilidad de Devolver el dinero de una venta

$$F(0,900) = P\left(X \le 0,900\right) = 1 - e^{-0,900} = 0,5934$$

Lo cual significa que a la larga, aproximadamente en el 59% de las ventas debe devolverse el dinero al no cumplir el instrumento con la duración de más de 900 horas.

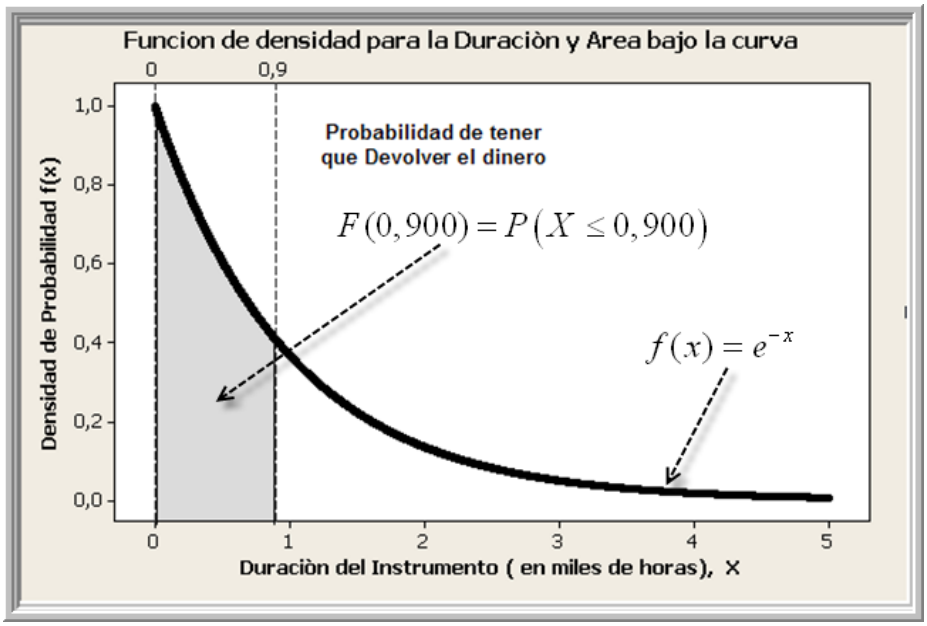


Figura 2.4M Representación de la probabilidad del evento "Devolver el Dinero"

c) *Calcule la distribución de probabilidad para la variable aleatoria "Utilidad de un Instrumento" (U)*

La variable aleatoria Utilidad U, tiene como espacio Muestral:

$\Omega_U = \{-\$2, \quad \$3\}$ es decir cuando le toca devolver el dinero, pierde los \$2 del costo y cuando no devuelve, gana \$3.

$$P(U = -\$2) = P(X \leq 0,900) = F(0,900) = 0,5934$$

$$P(U = \$3) = 1 - P(X \leq 0,900) = 1 - F(0,900) = 0,4066$$

Utilidad U	Probabilidad
-$ 2	0,5934
$ 3	0,4066
TOTAL	1,0000

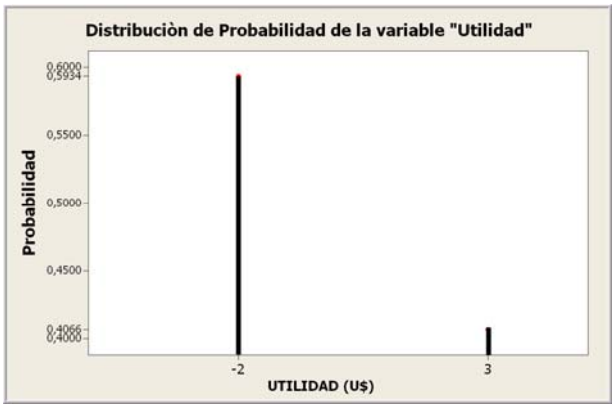


Figura 2.4N. Distribución de Probabilidad de la variable Aleatoria Utilidad en la venta de un Instrumento

2.3 CUARTILES DE UNA DISTRIBUCIÓN

Unas medidas cada vez mas utilizadas, son los cuartiles, que son tres valores Q_1, Q_2, Q_3 que dividen la muestra ordenada en cuatro partes que contienen aproximadamente el mismo numero de datos (de allí su nombre), es decir que el 25% de los datos son menores que Q_1, el 50% de los datos son menores que Q_2 y el 75% de los datos son menores que Q_3. Estos tres valores producen una muy buena síntesis de la distribución de frecuencias.

Nótese que siempre entre los valores Q_1 y Q_3, se encuentra el 50% central de los datos.

Calculemos los cuartiles para el ejemplo anterior, del tiempo de espera en un servicio de urgencias.

Primer cuartil Q_1

Note que el primer cuartil Q_1, se encuentra en el intervalo 11.15 a 13.15, puesto que la frecuencia acumulada hasta 11.15 es F(11.15)= 14% y F(13.15)=26%. Por lo tanto debe existir un punto Q_1, en dicho intervalo, tal que su frecuencia acumulada sea el 25%, es decir:

$$F(Q_1)=25\%.$$

Atendiendo a la expresión de F(x) para ese intervalo puede escribirse:

$$0.25 = F(Q_1) = 0.14 + \frac{0.12}{2}(Q_1 - 11.15)$$

De donde puede despejarse Q_1, obteniéndose el primer cuartil $Q_1 = 12.98$ minutos. Es decir que el 25% de las personas son atendidas en 12.98 minutos o menos.

Segundo cuartil Q_2. (Mediana)

Se desea encontrar el tiempo Q_2, tal que el 50% de las personas son atendidas en ese tiempo o menos, es decir: $F(Q_2) = 50\%$.

Al observar el cuadro o la función $F(x)$, encontramos que $F(13.15)=26\%$ Y $F(16.15)=56\%$, lo cual nos indica que el segundo cuartil Q_2, se encuentra entre 13.15 y 16.15, Si revisamos la función $F(x)$ para este intervalo y reemplazamos x por Q_2, se obtienen

$$F(Q2) = 0.26 + \frac{0.30}{3}(Q2 - 13.15) = 0.50$$

Despejando Q_2, resulta $Q_2 = 15.55$ minutos. Es decir que la mitad de la gente (50%), espera 15.55 minutos o menos.

Tercer cuartil

Siguiendo el proceso anterior, para $F(Q_3) = 75\%$, se obtiene que

$$F(Q3) = 0.74 + \frac{0.16}{3}(Q3 - 18.15) = 0.75$$

De donde al despejar resulta $Q_3 = 18.35$ minutos. Lo cual se interpreta como que el 75% de las personas esperan 18.35 minutos o menos.

Los cuartiles proporcionan una muy buena idea de la forma como están distribuidos los datos, pues entre un par de cuartiles consecutivos siempre está el 25% de los datos. Esta interpretación de la información que se obtiene de los cuartiles se hará más evidente en los diagramas de cajas y alambres, que se presentará más adelante.

Nótese que entre los cuartiles Q_1 y Q_3 siempre se encuentra el 50% central de los datos, pues abajo de Q_1 esta el 25% y arriba de Q_3 esta el 25%.

En el ejemplo anterior diríamos que la mitad de las personas esperan entre 12.98 y 18.35 minutos.

A la distancia entre los cuartiles Q_1 y Q_3, se le llama rango intercuartílico.

Rango intercuartílico (RIC) $= Q_3 - Q_1$. Para el ejemplo tendríamos que RIC= 5.37 minutos

2.3.1 Diagrama de caja y Alambres[15]

Este diagrama constituye una síntesis muy buena de la distribución de frecuencias y su sencillez la hace más útil, sobre todo en aquellas situaciones donde se hace necesario comparar dos o más distribuciones (poblaciones o tratamientos).

En la figura, se ilustra un diagrama de caja y alambres para el caso del ejemplo de los tiempos de espera.

Veamos cómo fue construido y cuál es su interpretación.

Se calculan los siguientes puntos:

$Q_1, Q_2, Q_3, \quad Q_1 - 1.5\,RIC, \quad Q_3 + 1.5RIC.$

[15] Estos gráficos son una contribución del gran estadístico Jhon Tukey.

: Q_1-1.5RIC =12.98 - 1.5(5.37) = 4.92 A este punto se le conoce como: "cerco interno inferior"

: Q_1 = 12.98 (primer cuartil)

: Q_2 = 15.55 (segundo cuartil = mediana)

: Q_3 = 18.35 (tercer cuartil)

: Q_3 + 1.5RIC = 18.35 + 1.5(5.37) = 26.40 "cerco interno superior".

Entre los cercos interiores, generalmente se encuentra un porcentaje alto de los datos, de tal manera que los puntos que se salen de los cercos, son puntos sospechosos de ser "OUTLIERS[16]" (Puntos atípicos).

---

[16] Los datos caracterizados como OUTLIERS tienen gran importancia, pues son puntos que tienen magnitudes "raras" con respecto al conjunto de datos. Es muy importante señalar que lo "raro", supone un criterio de lo que es "normal", de tal manera que se supedita a esa definición. Un punto puede ser raro, si se supone que la distribución de la cual proviene es Gaussiana (campana de Gauss), pero puede no serlo si su población de origen es una Weibull (forma de bañera). El señalar algunos puntos como OUTLIERS obliga a poner especial atención sobre ellos, puede ser desde una mala medición, hasta un verdadero hallazgo. En no pocas ocasiones los OUTLIERS se convierten en los puntos mas valiosos de una investigación. Imagínese un perno con una resistencia extraordinariamente superior a lo corriente.

Cuando se verifica que el dato es válido (medición correcta), en necesario definir la manera de involucrarlo en los análisis (ponderación). Un libro que trata de estos aspectos es BARNETT and LEWIS. "Outliers in Statistical data".

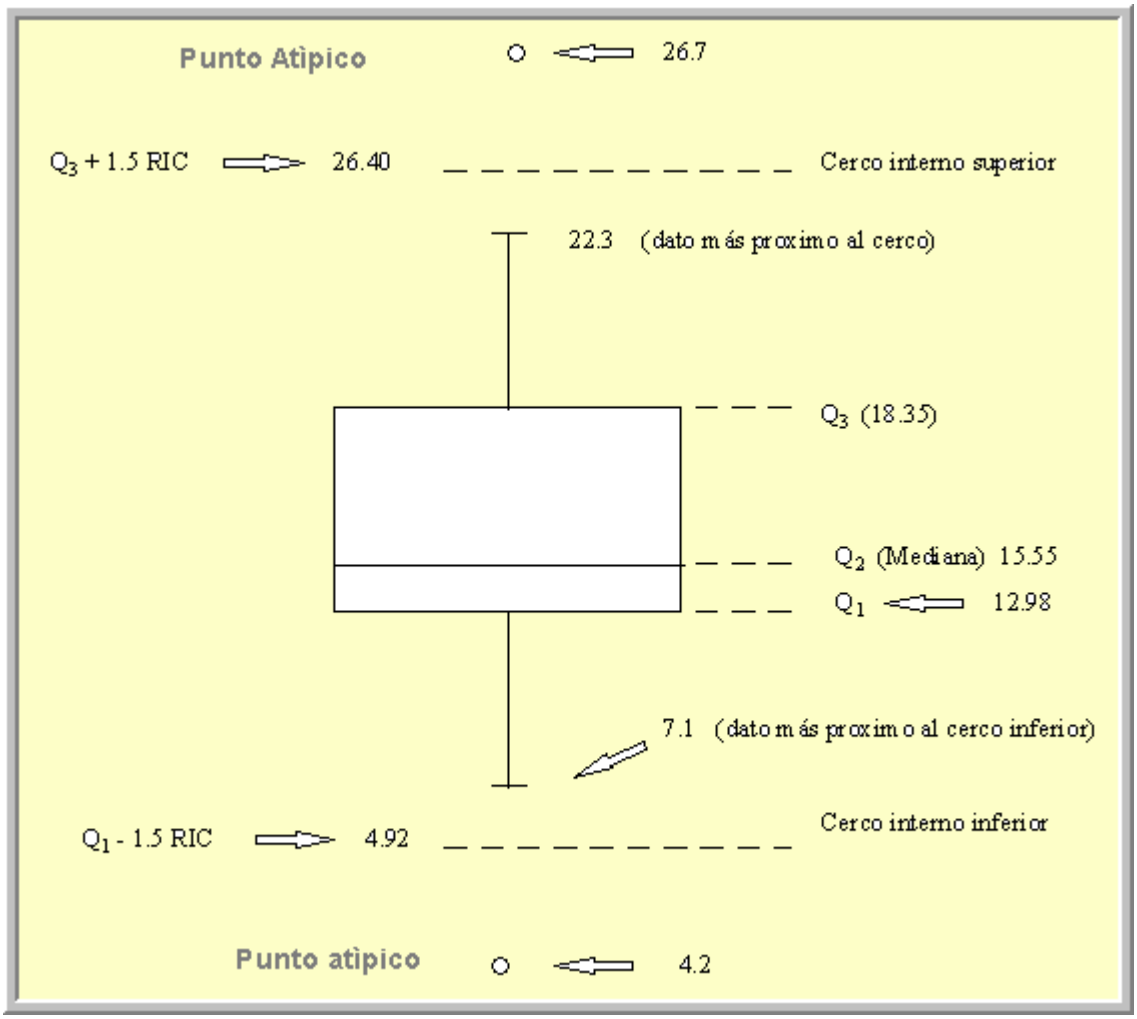


Fig. 2.5. Diagrama de caja y alambres para la distribución de los tiempos de espera en el servicio de urgencias de un hospital.

Con esta información se procede así: la caja se construye entre los cuartiles Q_1 y Q_3, con un ancho arbitrario. Dentro de la caja se marca Q_2, con trazo. Los alambres que salen de Q_1 y Q_3, van hasta el dato más próximo al cerco interno (sin cruzar el cerco.). Note que en este caso dichos puntos son 10.2 (que es el dato mas próximo al cerco interno inferior, que esta en 4.92) y por arriba esta el punto 22.3 (El dato mas próximo al cerco interno superior que es 26.4). Los puntos que se salen del cerco son marcados sobre el gráfico.

Se marcan (dibujan) los puntos que se han salido del cerco, en este caso son: 4.2 por abajo y el dato 26.7 que se salió del cerco interno superior.

También suele definirse un "cerco externo" ubicado a 3RIC de Q1 y Q3. Los puntos que quedan fuera de este cerco externo se conocen como OUTLIERS y son puntos que pueden ser atípicos, comparados con el cuerpo de datos. (En nuestro caso el cerco externo estaría entre los puntos -3.13 y 34.46, fuera de los cuales no se encuentra ningún dato.)

2.3.2 Como calcular los cuartiles, cuando los datos no están agrupados

Ejemplo 2.3

Los siguientes datos corresponden a las edades de 14 personas seleccionadas al azar, entre cierta clase de empleados de la población objetivo de un estudio.

25, 38, 29, 42, 39, 54, 23, 33, 45, 45, 26, 34, 30, 31.

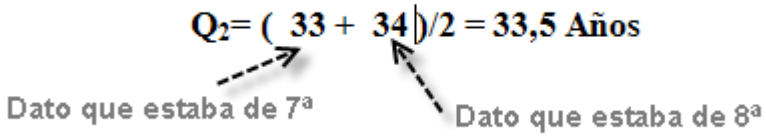
Pasó #1; Ordenar los datos de menor a mayor:

Observe que cuando los números indican "posición", los colocamos entre paréntesis.

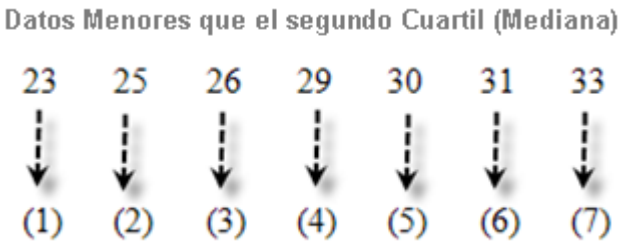
Los cuartiles los descubrimos calculando **la posición** que ocupan; es conveniente empezar por el segundo cuartil

Segundo cuartil Q_2**. (Mediana)**

Para calcular la posición que ocupa el segundo cuartil, promediamos las posiciones extremas ocupa la posición (14)+(1) / 2 = (7.5). Como existe la posición 7.5, porque un dato queda en la posición 7ª o en la 8ª, entonces que interpretaremos que queda en el medio de los datos que están de 7º y 8º , para evitar esta riña, hacemos el promedio de los dos datos que ocupan esas posiciones:



Primer Cuartil[17]**,** Q_1. El primer cuartil se obtiene considerando solo los datos que quedan antes de la mediana. Para este grupo de datos se calcula la media .Se trata pues de encontrar la posición de **la mitad de la mitad.**



La **posición** que ocupara el primer cuartil será la mediana de este primer grupo de datos: que es el que ocupe la posición

(7) +(1)/2 = (4.)

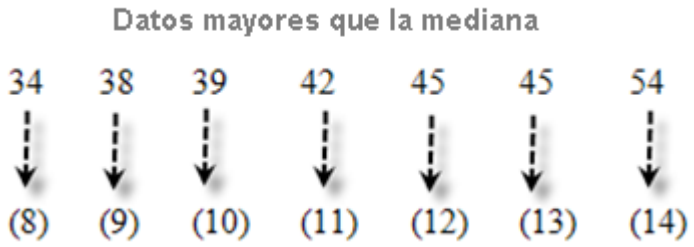
[17] Note que si el número de datos es impar, el segundo cuartil Q_2, resultaría ser un dato de la muestra. En este caso, para calcular la ubicación del primer cuartil Q_1, se toman en cuenta los datos que quedaron antes del segundo cuartil, excluyendo el dato que resulto ser el segundo cuartil Q_2. Análogamente para el tercer cuartil Q_3.

La Cuarta posición la ocupa el dato 29. Este es el primer cuartil.

Es decir que el **primer cuartil, Q_1** es el dato que ocupa la 4º posición, o sea que **Q_1 = 29 Años**

Si aplicamos este mismo procedimiento a los datos mayores que la mediana, se obtiene el tercer cuartil

El tercer cuartil Q_3.



La **posición** que ocupara el tercer cuartil será la mediana de este segundogrupo de datos: que es el que ocupe la posición

$$(8) +(14)/2 = (11.)$$

La posición once la ocupa el dato 42. Este es el tercer cuartil.

Q3 = 42 Años

Para la construcción de un diagrama de caja y alambres, se requiere de algunos cálculos adicionales, basados en los cuartiles ya encontrados:

RANGO INTERCUARTILICO (RIC)

RIC = Q_3-Q_1 = 42-29= 13 Años

EDAD MINIMA = 23 Años

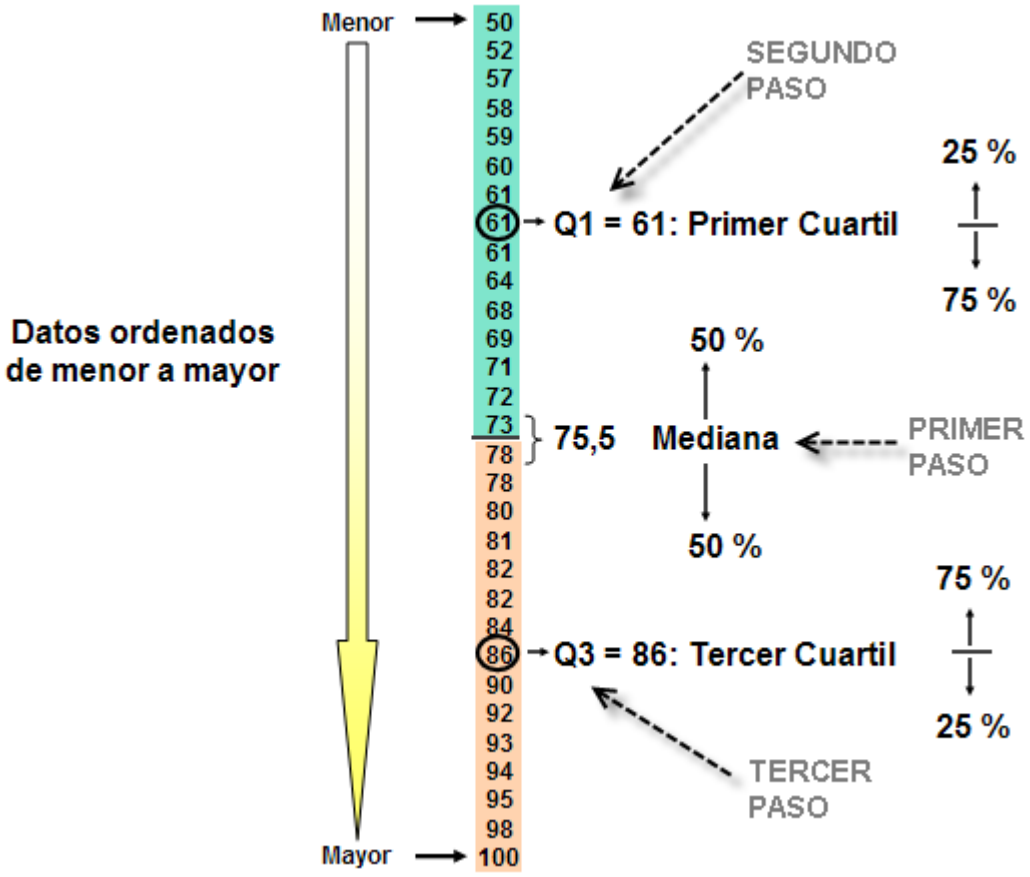
EDAD MAXIMA = 54 Años

cerco interno inferior = Q1- 1.5(RIC) = 29-1.5(13) = 9.5

cerco interno superior = Q3 + 1.5(RIC) = 42 + 1.5(13)= 61.5

Construya usted el diagrama para este caso[18].

Otro ejemplo (Sìntesis)



[18] Note que en este caso particular, todos los puntos quedaron dentro de los dos (2) cercos, lo cual no ocurre siempre, por esta razón los puntos interiores mas cercanos al cerco son el mínimo y el máximo de los datos, que definen la longitud de los "alambres" que van pegados a la caja.

En resumen puede decirse que los diagramas de cajas y alambres son útiles, entre otros para los siguientes propósitos:

1. Para identificar la localización de los datos alrededor de la mediana.

2. Para hacerse una muy buena idea de la dispersión de los datos, basándose en la longitud de la caja (rango intercuartílico), pues siempre la caja, corresponde al 50% de los datos que están en la parte central. Además se aprecia el rango de los datos, el cual corresponde a la distancia entre las observaciones más extremas.

3. El diagrama de cajas y alambres, nos permite hacernos una muy buena idea sobre el grado de asimetría de una distribución, al comparar la proporción de la caja que queda a la izquierda de la mediana, con la que queda a la derecha, igualmente la longitud de los alambres respectivos. En el ejemplo de la figura, se observa que los datos estan más concentrados en entre Q_1 y Q_2 que entre Q_2 y Q_3, lo cual es una muestra de cierto grado de asimetría.

4. El diagrama es útil para identificar posibles OUTLIERS (fuera de los cercos internos pero dentro de los externos) y OUTLIERS (fuera de los cercos externos).

5. Una utilidad grande de los diagramas de caja y alambres, es comparar varias poblaciones, a través de sus distribuciones. En este caso se construye un diagrama para cada distribución y se dibujan en una misma escala (sobre un mismo plano), lo cual permite muy fácilmente hacerse una idea de las semejanzas y las diferencias de los rasgos más importantes de las distribuciones. Como se ilustrara en un ejemplo más adelante.

Ejemplo 2.4

En el cultivo de la caña de azúcar, se llama una "suerte" a un lote de terreno, en el cual hay varias parcelas del cultivo, a las cuales se les da el mismo tratamiento, es decir cuando se cosecha, se hace en todas las parcelas de la suerte, cuando se arregla el terreno igualmente o cuando se siembra o se riega. El terreno de una suerte puede llegar a ser usado hasta para cuatro

siembras consecutivas antes de ser "acondicionado" de nuevo (remover tierra, agregar abono, fertilizantes, etc.). Se supone que con cada siembra el terreno se fatiga y que eso se verá reflejado en la producción de caña (o en la de azúcar).

Se han tomado datos de producción de varias suertes, que han estado sometidas a diferente número de cortes (o de siembras), que tienen diferente procedencia (caña propia (1) o de proveedor externo (0), edad de corte (meses). Use un diagrama de cajas para comparar las distribuciones de frecuencias de los rendimientos para las suertes de acuerdo con los diferentes criterios, que se menciona en el problema.

PREGUNTA 1: El número de cortes que se haya hecho sobre un terreno, desde su último acondicionamiento, afecta el rendimiento?

Para dar respuesta a esta pregunta, debe compararse las distribuciones del rendimiento para las poblaciones que tienen distinto número de cortes. A continuación se comparan, a través de diagramas de cajas.

Se puede observar en la figura 2.6 en forma contundente que el número de cortes afecta considerablemente el rendimiento, note por ejemplo que la caña sembrada en un terreno con cuatro cortes, tiene un rendimiento mediano de alrededor de 83 Ton/Fa, mientras la de tres (3) cortes tiene alrededor de 110 Ton/Fa, la de dos (2) cortes 130 Ton/fa y la de un corte tiene un rendimiento mediano de aproximadamente 143 Ton/fa.

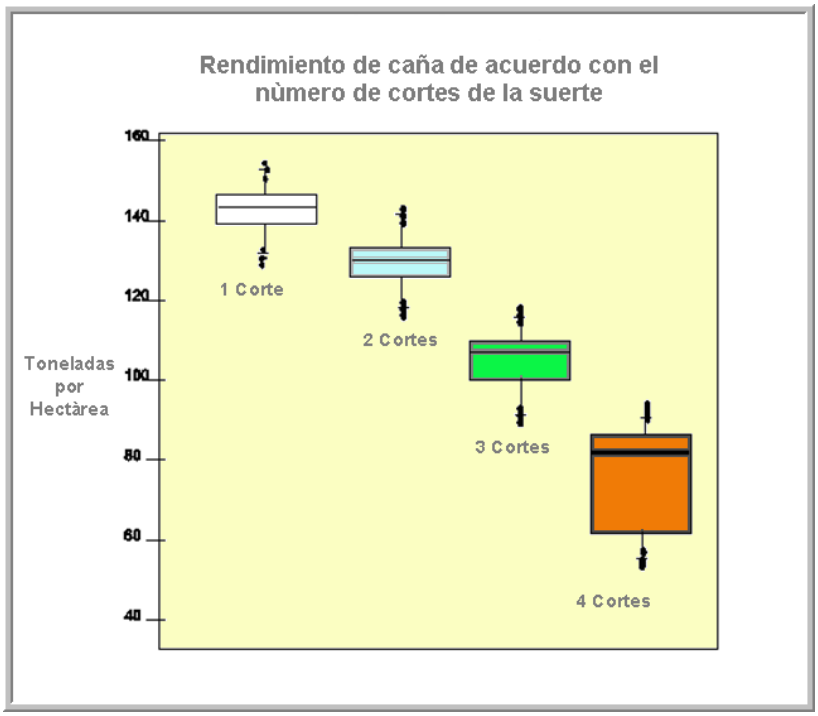


Fig. 2.6. Diagrama de cajas

En la Figura 2.7, puede notarse que las distribuciones, para los cortes 1, 2, 3 tienen variabilidad muy parecida, mientras que la variabilidad de la distribución del rendimiento para las de cuatro (4) cortes es mayor.

Nótese también que en esta situación se han considerado en forma conjunta la producción propia del ingenio y la de los proveedores externos, por eso surge de manera natural la pregunta siguiente.

PREGUNTA 2. El comportamiento registrado en la anterior situación, es válida independientemente de si el origen de la caña es "ingenio" o "proveedor"?

Para dar respuesta a esta pregunta, deben construirse los diagramas de caja para cada número de cortes, separadamente para caña del "ingenio" y para "proveedores", como se muestra en la

figura 2.7. De esta manera estamos valorando la "procedencia" como un posible factor de confusión.

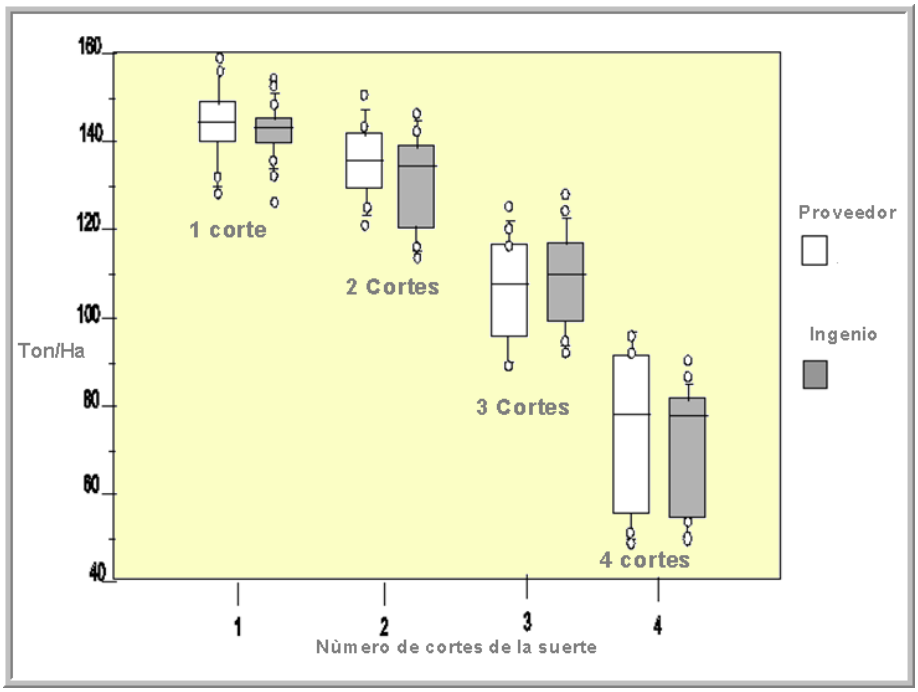


Fig. 2.7. Diagrama de cajas de la comparación del rendimiento de acuerdo con el origen de la caña y el número de cortes en la suerte

Observe en la gráfica las cajas sombreadas corresponden a las distribuciones del rendimiento, para caña del "ingenio", mientras la blanca corresponde a "proveedor" externo. Se nota un comportamiento bastante similar, es decir, no parece existir diferencia en la caña con respecto a su origen. Los rendimientos medianos, son consistentes con los del primer gráfico, al igual que su variabilidad.

La edad de corte, parece tener bastante importancia, averigüemos ahora por su distribución:

PREGUNTA 3: Cuál es la distribución de la edad de corte, de acuerdo con el origen de la caña y de su número de cortes?

Para dar respuesta a este interrogante, se construyen cajas para la variable "edad de corte" (meses), para cada una de las distintas subpoblaciones que resultan de la combinación de número de cortes y origen (procedencia).

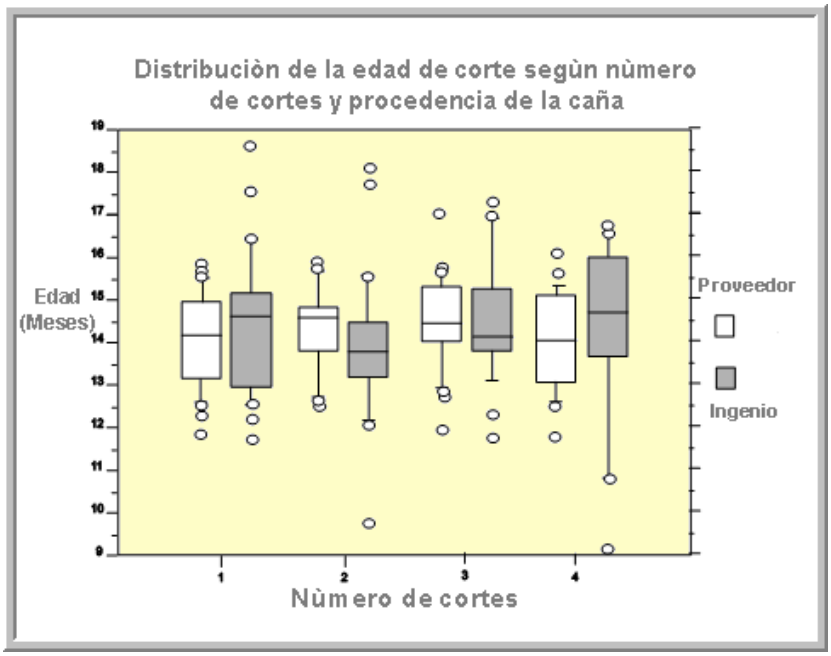


Fig. 2.8. Comparación de la edad de Corte según el numero de cortes que se han practicado en la suerte

En esta situación, sería muy conveniente conocer un poco más sobre el fenómeno, para tener claridad acerca de cuál es la edad óptima de corte, aunque depende de la variedad de caña que se siembre. Supongamos que para nuestro caso, la edad de corte recomendada está entre 12.5 y 13.5 meses. A medida que la caña envejece va empobreciendo su contenido de sacarosa, que es en realidad lo que interesa. En estas condiciones podría decirse que en casi todos los casos se corta después de 12,5 meses, sin embargo, un porcentaje muy grande de las veces se está cortando por encima de los 13.5 meses. Se sugiere averiguar las razones para que esto esté ocurriendo.

PREGUNTA 4 .¿Cómo afecta la edad de corte, el rendimiento de la caña en cuanto al volumen de caña cosechado? (Note que aquí no sabremos el impacto en términos del contenido de sacarosa, solo del rendimiento en términos de la cantidad de caña colectada).

Dado que ya conocemos que el número de cortes, es una variable importante, debemos involucrarla en el análisis, para que no se convierta en un factor de confusión. De esta manera debe construirse las cajas para la distribución del rendimiento, para cada categoría de número de cortes y de edad. Aquí, la edad se ha categorizado, en tres grupos: joven, madura y vieja. Veamos el resultado.

Obsérvese en la figura 2.9, que para cada número de cortes hay tres gráficos que corresponden a diferentes grados de madurez de la caña al cortarse, pero sistemáticamente, en cada uno de los grupos de tres gráficos, la distribución de la caña joven, tiene un rendimiento mediano mas alto, seguido por la madura y por último por la vieja, presentándose diferencias relativamente mas grandes en la caña de cuatro (4) cortes.

En esta comparación se ve muy claro el impacto de la edad de corte.

Queda pendiente un estudio, en el que se evalúe el contenido de sacarosa y podría repetirse el análisis, teniendo como variable de respuesta Ton de azúcar/Fa.

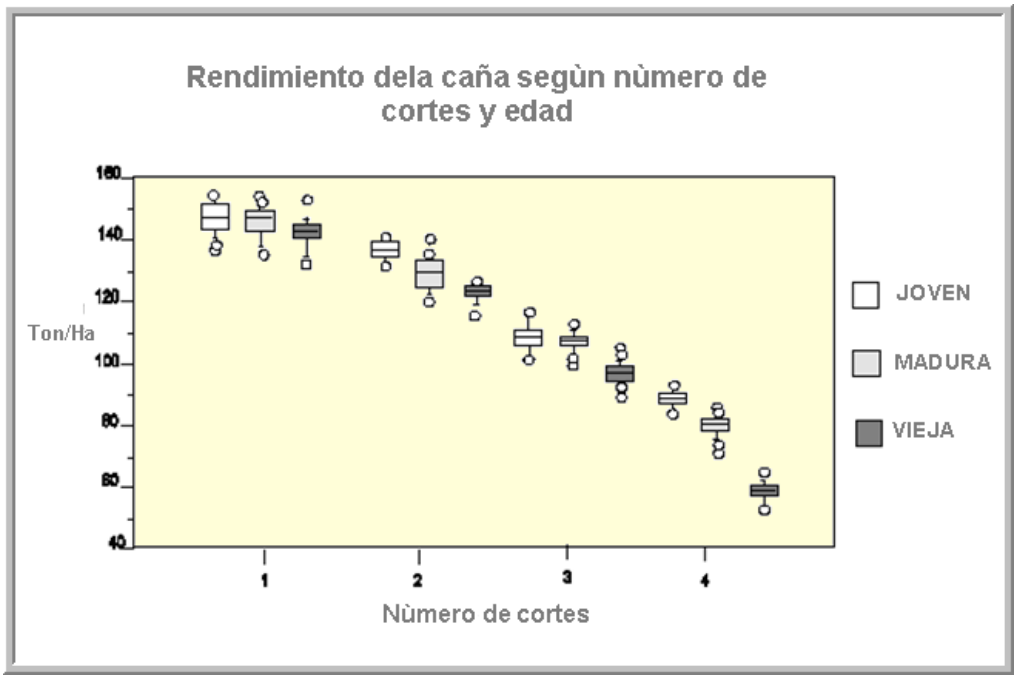


Fig. 2.9. Distribución de la Edad de corte según numero de cortes se la suerte.

2.4 REDUCCION DE DATOS

Hasta ahora se ha tratado de organizar la información, resumiéndola a través de los cuadros de frecuencias y de la representación gráfica, no obstante en ocasiones se requiere de algunas medidas que en forma muy directa puedan indicar rasgos importantes de la muestra, como su magnitud, su homogeneidad, su simetría, etc. Al proceso de resumir los datos por medio de estadígrafos que indiquen sus rasgos, se denomina **reducción de datos.**

Se comenzará con la presentación de algunos indicadores de la magnitud, de los datos de la muestra que han sido llamados:

2.4.1 Indicadores de tendencia central

Entre los principales indicadores se consideran los siguientes:

Media aritmética, mediana, moda, y media geométrica.

2.4.1.1 La media aritmética

La media aritmética de una muestra de datos: $x_1, x_2,..., xn$, se define como:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

Si los datos corresponden a una variable discreta que está organizada en un cuadro de frecuencias, se puede escribir:

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i x_i}{n} = \sum_{i=1}^{m} \frac{n_i}{n} \times x_i = \sum_{i=1}^{m} f_i x_i$$

Ejemplo 2.5

Sean 2, 3, 2, 2, 2, 3, 1, 3, 3, 4, una muestra de tamaño $n = 10;$ su media aritmética será:

$$\bar{x} = \frac{2 + 3 + 2 + 2 + 2 + 3 + 1 + 3 + 3 + 4}{10} = 2.5$$

Si la muestra se presenta en un cuadro de frecuencias tenemos:

x_i	n_i	f_i
1	1	0.1
2	4	0.4
3	4	0.4
4	1	0.1

y la media puede calcularse como:

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i x_i}{n} = \frac{1\times 1 + 4\times 2 + 3\times 4 + 1\times 4}{10} = 2.5$$

Ó lo que es exactamente lo mismo como

$$\bar{x} = \sum_{i=1}^{m} f_i x_i = \ 0.1 \text{ x } 1 + 0.4 \text{ x } 2 + 0.4 \text{ x } 3 + 0.1 \text{ x } 4 \ = \ 2.5$$

Propiedades de la media aritmética

1. La suma de las desviaciones de los datos con respecto a la media es cero.

   * definimos desviación del dato x_i con respecto al valor *"a"* como:

$$d_i \ = \ x_i \ - \ a$$

Así que la propiedad puede escribirse como:

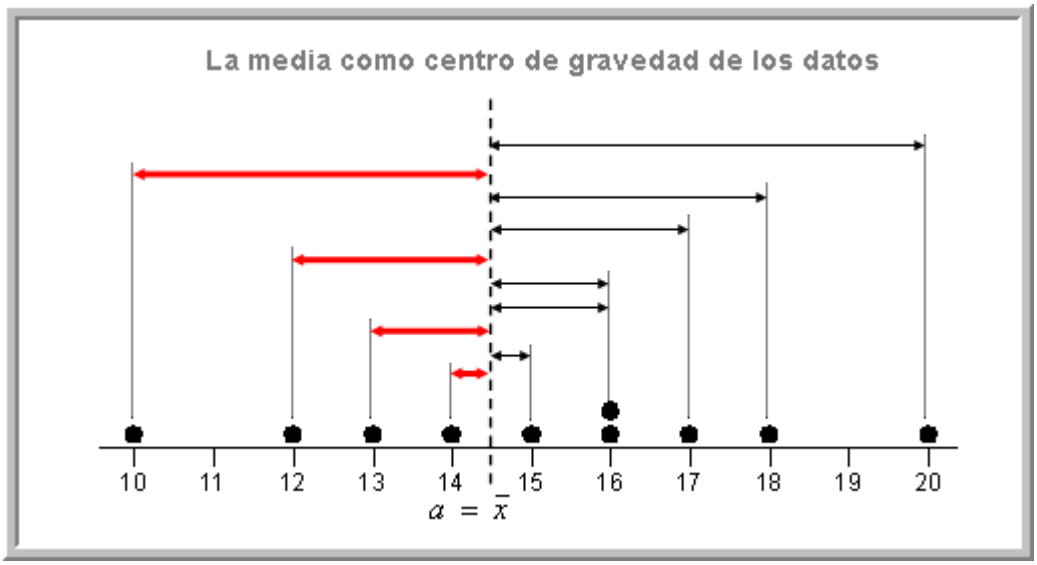
$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

La verificación puede hacerse en forma sencilla:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x}$$

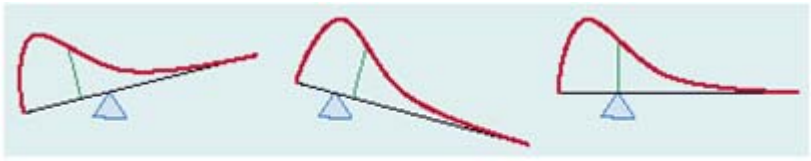
$$= \sum_{i=1}^{n} x_i - n\left(\frac{\sum x_i}{n}\right) = 0$$

Esta propiedad refuerza la media como indicador de tendencia central. Su significado es el siguiente:



Como puede apreciarse, las desviaciones de los datos que están a la izquierda de la media tienen signo negativo y las de la derecha signo positivo, por esta razón, para que la suma de todas sea cero, debe suceder que la suma de las distancias a la media de los datos de la izquierda de ella, debe ser igual a la suma de las distancias a la media de los datos de la derecha, lo cual convierte a la media en el **centro de gravedad.**

Si quisiéramos visualizar esta propiedad a partir de una distribución expresada en términos de su función densidad :



La media como centro de gravedad

La interpretación física nos dice que si justo donde se ubica la media aritmética se colocara un punto de apoyo y se colgara de los puntos donde se ubican los datos, el mismo peso en cada uno, entonces el sistema quedaría en equilibrio.

2. La media de los cuadrados de las desviaciones de los datos con respecto a un valor *"a"* es mínima, cuando $a = \bar{x}$. Es decir:

$$\text{f(a)}=\frac{\sum_{i=1}^{n}(x_i-a)^2}{n} \text{ tiene su mínimo en a}=\bar{x}$$

Demostración:

$$\text{f(a)}=\frac{1}{\text{n}}\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)+\left(\bar{x}-a\right)\right]^2$$

Desarrollando el cuadrado:

$$\text{f}\left(\text{a}\right)=\frac{1}{\text{n}}\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)^2+2\left(x_i-\bar{x}\right)\left(\bar{x}-a\right)+\left(\bar{x}-a\right)^2\right]$$

$$=\frac{1}{\text{n}}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+2\frac{1}{\text{n}}\left(\bar{x}-a\right)\sum_{i=1}^{n}\left(x_i-\bar{x}\right)+\frac{n\left(\bar{x}-a\right)^2}{n}$$

Como $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=0$ (propiedad 1)

Entonces:

$$\text{f}\left(\text{a}\right)=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\left(\bar{x}-a\right)^2}{n}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n}+\left(\bar{x}-a\right)^2$$

Como puede apreciarse el primer término no depende de *"a"* y además $n(\bar{x}-a)^2 \geq 0$, por tanto f(a) es mínimo cuando $n(\bar{x}-a)^2 = 0$ y esto ocurre cuando $a = \bar{x}$.

3. Si $x_i = k$, para todo *i,* o sea que si todos los datos son iguales a *k*, entonces: $\bar{x} = k$.

Veamos:

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i}{n} = \frac{\sum_{i=1}^{n} k}{n} = \frac{nk}{n} = k$$

4. Si todos los datos de una muestra se multiplican por una constante, el promedio de dicha muestra resulta multiplicando por la misma constante, es decir:

si $y_i = ax_i$, $i = 1, 2, ..., n;$ entonces $\bar{y} = a\bar{x}$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} ax_i}{n} = a\frac{\sum_{i=1}^{n} x_i}{n} = a\bar{x}$$

5. Si $Z_i = ax_i + by_i$, $i = 1, 2, ..., n$; donde a, b son constantes, entonces

$$\bar{Z} = a\bar{x} + b\bar{y}$$

Veamos:

$$\bar{Z} = \frac{\sum_{i=1}^{n} Z_i}{n} = \frac{\sum_{i=1}^{n} (ax_i + by_i)}{n} = a\frac{\sum x_i}{n} + b\frac{\sum y_i}{n}$$

$$\bar{Z} = a\bar{x} + b\bar{y}$$

Esta propiedad puede generalizarse a la combinación lineal de k variables y puede resumirse diciendo que la media aritmética es un operador lineal.

Ejemplo 2.6

Se ha tomado una muestra de parejas de casados y se han observado las variables X e Y.

X : Ingreso mensual del esposo

Y : Ingreso mensual de la esposa

Se encontró que el ingreso promedio mensual de los esposos es

$\overline{X}$ = \$100.000 y de las esposas $\overline{Y}$ = \$80.000.

Si se define la variable ingreso familiar *Z*, como la suma de los ingresos de los esposos, entonces el ingreso familiar de la pareja *i* será: $Z_i = X_i + Y_i$ y el ingreso familiar promedio será:

$$\overline{Z} = \overline{X} + \overline{Y} = \$100.000 + \$80.000 = \$180.000$$

6. Si una muestra de *n* elementos, se divide en *k* submuestras excluyentes y exhaustivas, que tienen $n_1, n_2, ..., nk,$ elementos $(n_1 + n_2 + ... + n_k = n)$, con promedios $\overline{x}_1, \overline{x}_2, ..., \overline{x}_k$ respectivamente, entonces el promedio de la muestra global estará dado por:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + ... + n_k\overline{x}_k}{n}$$

es decir:
$$\overline{x} = \frac{\sum_{i=1}^{k} n_i\overline{x}_i}{n}$$

El promedio $\overline{x}_i$, de los datos del grupo *i*, está dado por: $\overline{x}_i = \frac{\sum_{G_i} x_j}{n_i}$

por tanto: $\sum_{G_i} x_j = n_i\overline{x}_i$

Por otro lado:

$$\sum_{j=1}^{n} x_j = \sum_{G_1} x_j + \sum_{G_2} x_j + ... + \sum_{G_k} x_j$$
$$= n_1\overline{x}_1 + n_2\overline{x}_2 + ... + n_k\overline{x}_k$$

Entonces: $$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n}$$

Ejemplo 2.7

Una muestra de 500 trabajadores tienen un salario promedio de $108.000, si el salario promedio de los hombres es $120.000, y el de las mujeres $100.000, ¿cuántos hombres y mujeres hay?

Si n_1 es el número de hombres y n_2 el de mujeres, entonces:

$$n_1 + n_2 = 500 \qquad (1)$$

Además:

$$\$108.000 = \frac{n_1 \times 120.000 + n_2 \times 100.000}{500} \qquad (2)$$

Resolviendo (1) y (2) se obtiene: $n_1 = 200$ y $n_2 = 300$

Cálculo de la media aritmética para los datos agrupados en intervalos de clase.

Se sabe que cuando los datos están agrupados en clases, se pierde la individualidad de la información, así por ejemplo puede conocerse que en el intervalo (*10,20*] hay 3 datos, pero no conocemos cuál es el valor de cada uno de estos datos; esto plantea una dificultad para el cálculo de la media usando la definición presentada.

Se puede calcular en este caso la media, en forma aproximada, usando la propiedad 6 y el supuesto de que los datos en cada intervalo están uniformemente distribuidos, puesto que si esto sucede , la media aritmética de los datos del intervalo *i,* coincide con el punto medio del intervalo (marca de clase), de esta manera se puede considerar la muestra total, dividida en *"m"* submuestras constituidas por los datos que pertenecen a cada uno de los intervalos, así aplicando la propiedad 6, se obtiene que:

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_m\bar{x}_m}{n}$$

Como: $\bar{x} \equiv x_i'$; entonces :

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i x_i'}{n} = \sum_{i=1}^{m} h_i \times x_i'$$

Ejemplo 2.8

Dada la siguiente distribución de frecuencias:

Intervalo de clase	Marca de clase	Frecuencia absoluta (n_i)	Frecuencia relativa (f_i)
(10,20]	15	12	0.12
(20,40]	30	16	0.16
(40,60]	50	42	0.42
(60,70]	65	25	0.25
(70,100]	85	5	0.05

La media aritmética de esta distribución será:

$$\bar{x} = \frac{12 \times 15 + 16 \times 30 + 42 \times 50 + 25 \times 65 + 5 \times 85}{100} = 48.1$$

O en forma equivalente:

$$\bar{x} = 0.12 \text{ x } 15 + 0.16 \text{ x } 30 + 0.42 \text{ x } 50 + 0.25 \text{ x } 65 + 0.05 \text{ x } 85$$

$$\bar{x} = 48.1$$

2.4.1.2 La mediana (Me)

La mediana ya fue tratada cuando tocamos el tema de los cuartiles, pues la mediana corresponde con el segundo cuartil. En síntesis la definimos de la siguiente manera.

Si $X_1, X_2, ..., X_n$ corresponde a una muestra de realizaciones (datos) de una variable X y ordenamos dichos valores de la forma: $X_{(1)}, X_{(2)}, ..., X_{(n)}$. Ahora hemos colocado los subíndices entre paréntesis para indicar las nuevas posiciones de los datos, es decir que el menor de los datos ahora se llama $X_{(1)}$ van en secuencia no decreciente, hasta llegar a $X_{(n)}$ que es el mayor de todos. Así las cosas la mediana se halla con la siguiente expresión:

$$M_e = \begin{cases} X_{\frac{n+1}{2}} & n \text{ impar} \\ \dfrac{X_{\frac{n}{2}} + X_{\left(\frac{n}{2}+1\right)}}{2} & n \text{ par} \end{cases}$$

Si quisiéramos definir la mediana con solo palabras, deberíamos decir que es un valor M_e, tal que supera no más de la mitad de los datos y es superado por no más de la mitad de los datos. (parece un trabalenguas, pero es una definición válida) A continuación se presentan algunos ejemplos: supóngase que se tiene la siguiente muestra ordenada en forma no decreciente: 2, 5, 7, 9, 11, veamos si 5 cumple la definición: 5 supera un dato (no más de la mitad de los datos) y es superado por 3 datos (más de la mitad), esto implica que 5 no es la mediana.

Probemos con el 7; éste supera dos datos (no más de la mitad) y es superado por dos datos (no más de la mitad), así que $M_e = 7$, se puede intuir que siempre que el número de datos sea impar, al ordenar la muestra, existirá un valor único tal que supera y es superado por el mismo número de datos, éste será la mediana.

Cuando el número de datos es par por ejemplo, sea la muestra 2, 5, 7, 9, 11, 15, ordenada en forma no decreciente, al aplicar la definición al valor 7; éste supera a 2 datos (no más de la mitad) y es superado por 3 datos (no más de la mitad) esto implica que 7 es mediana.

Al ensayar con el valor 9; éste supera tres datos (no más de la mitad) y es superado por dos datos (no más de la mitad), es decir que también 9 es mediana; nótese además que cualquier punto de la recta real, que se encuentre entre 7 y 9, cumple con la definición, en estos casos cuándo el número de datos es par, se ha convenido definir la mediana como el promedio de los dos datos que son medianos así pues:

$$M_e = \frac{7 + 9}{2} = 8$$

Cálculo de la mediana cuando los datos están agrupados en intervalos de clases.

Supóngase que se tienen *m* intervalos: $(L_0 , L_1] , (L_1 , L_2] , ..., (L_{m-1} , L_m]$, la mediana es el punto cuya frecuencia absoluta acumulada es n/2 ó la relativa acumulada es 0.50, es decir la mediana es el valor *x* tal que:

$$N(x) = \frac{n}{2}$$

o en forma equivalente:

$$F(x) = 0.50$$

De acuerdo con esto el intervalo $(L_{i-1} , L_i]$ que contiene la mediana es tal que:

$$N(L_{i-1}) < \frac{n}{2} \quad y \quad N(L_i) \geq \frac{n}{2}$$

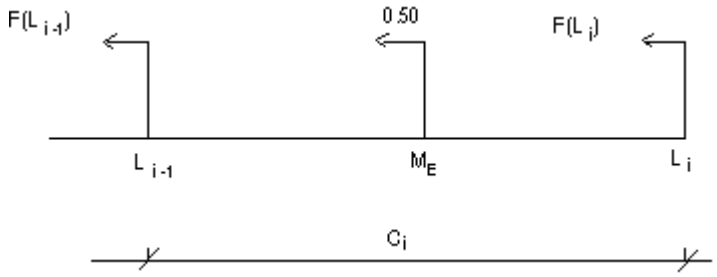
o lo que es equivalente:

$$F(L_{i-1}) < 0.50 \quad y \quad F(L_i) \geq 0.50$$

Una vez localizado el intervalo que contiene la mediana, se encuentra por interpolación el valor M_e, tal que:

$$M_e = N^{-1}\left(\frac{n}{2}\right) = F^{-1}(0.50)$$

y puede procederse con base en la definición de la función de distribución empírica vista anteriormente.



$$F(M_e) = 0.50 = F(L_{i-1}) + \frac{f_i}{C_i}(M_e - L_{i-1})$$

Despejando M_e de la anterior expresión tenemos:

$$M_e = L_{i-1} + \frac{0.50 - F(L_{i-1})}{f_i} * C_i \qquad (2.7)$$

donde f_i es la frecuencia relativa del intervalo de clase que contiene la mediana.

Expresada en términos de la frecuencia absoluta:

$$M_e = L_{i-1} + \frac{\frac{n}{2} - N(L_{i-1})}{n_i} * C_i \qquad (2.8)$$

Ejemplo 2.9

Intervalo de clase	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia acumulada	
			absoluta (N_i)	relativa (F_i)
(10,20]	6	0.12	6	0.12
(20,40]	8	0.16	14	0.28
(40,60]	21	0.42	35	0.70
(60,70]	13	0.26	48	0.96
(70,100]	2	0.04	50	1.00
	50	1.00		

Si se observan las frecuencias acumuladas puede notarse que el 28% de los datos son menores o iguales que 40 y que el 70% son menores que 60, lo cual implica que debe existir un punto en el intervalo (40 , 60] tal que el 50% de los datos sean menores o iguales que él; lo cual indica que el intervalo (40 , 60] contiene la mediana. De acuerdo con la expresión (2.7), se tiene que:

$L_{i-1} = 40$

$F(L_{i-1}) = 0.28$ $\qquad M_e = 40 + \dfrac{0.50 - 0.28}{0.42} \times 20 = 50.5$

$f_i = 0.42$

$C_i = 20$

Propiedad de la mediana

La suma de las distancias de los datos a un punto *"a"* es mínima cuando ese punto es la mediana, es decir:

Si f(a) $= \sum_{i=1}^{n} |x_i - a|$, entonces f(a) tiene un mínimo en

$a = M_e$.

Para una mejor interpretación de esta propiedad, se presenta el siguiente ejemplo:

Ejemplo 2.10

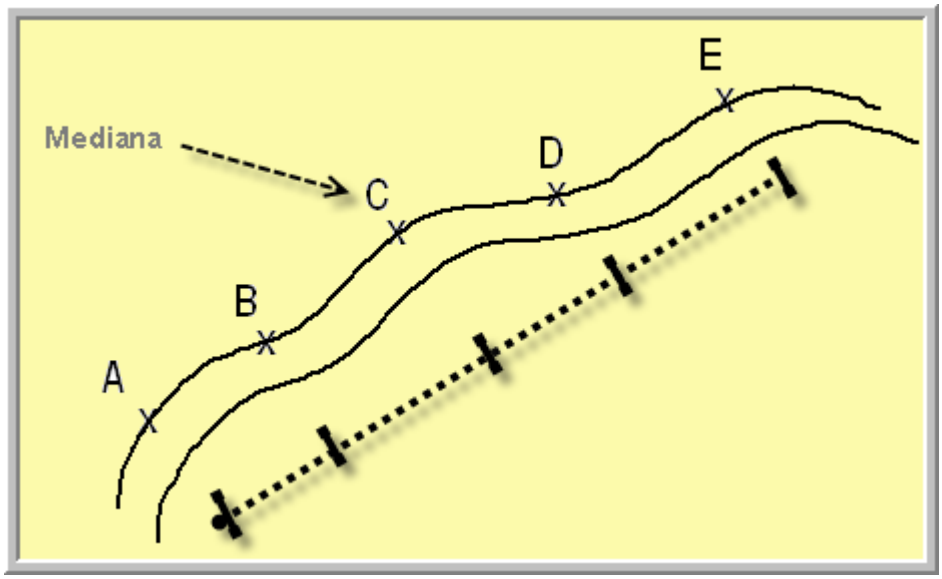


Figura 2.10.

En la figura 2.10 se muestra la posición relativa de las poblaciones *A, B, C, D y E*, si la demanda de todas las poblaciones por cierto tipo de artículo puede asumirse igual, ¿en cuál población debe colocarse la fábrica de dicho artículo si se quiere minimizar la distancia promedio a recorrer?

La respuesta a dicha pregunta puede darse mediante la siguiente reflexión: si se escoge un origen arbitrario sobre la carretera para medir los recorridos desde cada población a dicho origen, podremos notar que el recorrido a la población *C* es la mediana, lo cual significa de acuerdo con la propiedad que la suma de las distancias de las demás poblaciones hasta la población *C* es la mínima posible y por lo tanto su promedio también será mínimo, de esta manera la fábrica debe colocarse en la población *C* si se quieren minimizar los costos de transporte.

Otra propiedad de la mediana se explica a continuación:

La sensibilidad es una cualidad deseable en un indicador, puesto que ello implica qué cambios producidos en la muestra pueden ser detectados por el indicador; pero mucha sensibilidad en un

indicador puede ser contraproducente, puesto que cambios irrelevantes en la muestra pueden producir grandes cambios en el indicador, lo cual puede prestarse para interpretaciones equivocadas, esto ocurre con la media aritmética, cuando la distribución es asimétrica, es decir cuando hay unos pocos valores muy grandes o muy pequeños, la media es muy afectada por ellos.

Ejemplo 2.11

Si los salarios de los empleados de una empresa tienen la siguiente distribución:

Distribución de frecuencias del salario de la empresa "Break Sure"

Salario x_i	porcentaje de personas (f_i)
$ 100.000	20 %
$ 120.000	10 %
$ 130.000	25 %
$ 150.000	40 %
$1.200.000	5 %
	100 %

Si se pretende formar una idea de la magnitud de los salarios de dicha empresa, usando la media aritmética se tiene:

$$\bar{x} = \sum_{i=1}^{m} x_i f_i = \$10.000 \times 0.20 + \$12.000 \times 0.10 + \$3.000 \times 0.25 +$$
$$+\$15.000 \times 0.40 + \$120.000 \times 0.05$$
$$\bar{x} = \$18.450$$

Como puede apreciarse, un 5% de valores muy grandes influyen tanto en la media, que su valor $184.500, es superior al 95% de los salarios por esta razón, en este caso, la media aritmética, mal podría representar la muestra.

La mediana en cambio es más **resistente** a los valores extremos, en este caso, la mediana corresponde al valor $M_e = \$130.000$.

2.4.1.3 La moda

Cuando la variable de interés, es de naturaleza discreta, la moda $\boldsymbol{M_0}$ corresponde al dato de la muestra que tiene mayor frecuencia, por ejemplo, en la muestra: 2, 3, 1, 1, 1, 4, 3, 1, 5, 1, 5, 2, la moda es $M_0 = 1$ puesto que posee la mayor frecuencia (aparece 5 veces).

Cuando se trata de una variable de naturaleza continua, la moda corresponde al(os) valor(es) alrededor del(os) cual(es) se produce una mayor concentración de datos, es decir a los puntos de mayor densidad de frecuencia. En lenguaje matemático diríamos, refiriéndonos a la función de densidad de frecuencia o de probabilidad, que la(s) moda(s) corresponden a los cpuntos que son máximos locales, como muestra la figura 2.11.

Si se conociera la función de densidad poblacional (ver Fig.2.11) la moda corresponde a sus máximos relativos; en la función que muestra el gráfico se aprecian 3 modas.

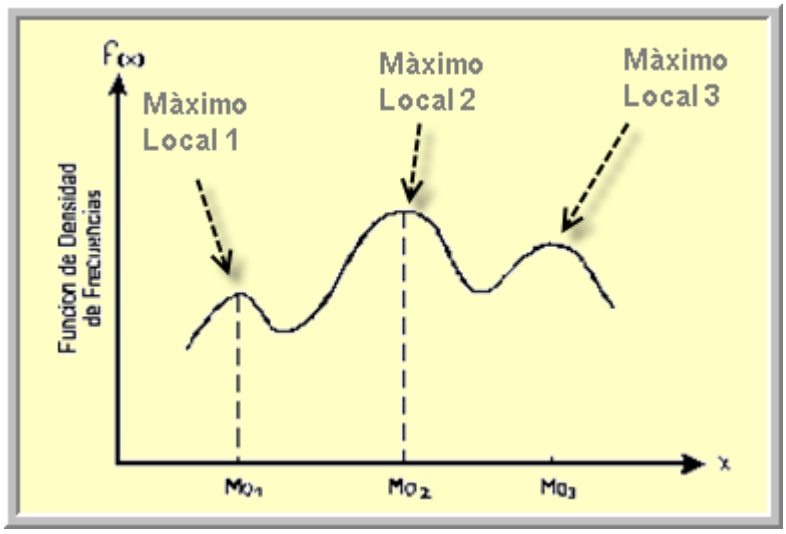


Fig. 2.11: Gráfico de la función de densidad de frecuencia poblacional de alguna variable *X*.

Los tres puntos que muestra la figura 2.11, son puntos de máxima densidad en sus entornos respectivos. Si se conociera la función de densidad en forma analítica, podríamos encontrar la(s) moda(s), por derivación de la función *f(x);* pero como sólo se dispone de la función de densidad empírica que se construyó a partir de la muestra, se debe definir un procedimiento que permita acercarse a la determinación de los mencionados máximos relativos, para ello se hace referencia a la figura 2.12.

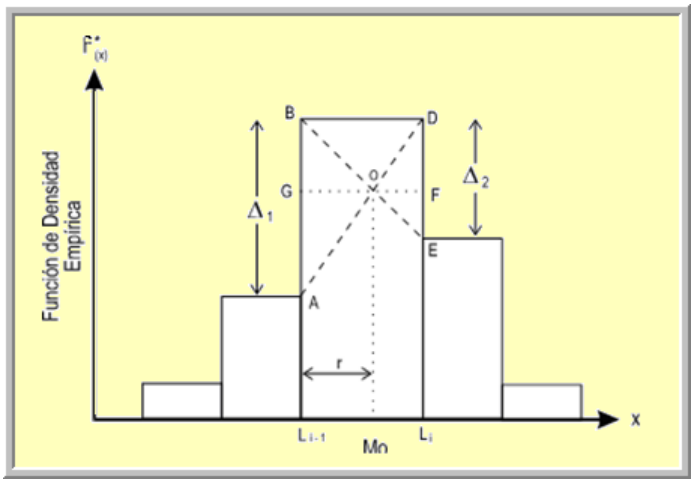


Fig. 2.12: Función empírica de densidad. Elementos que intervienen en le cálculo de la moda

Se supone que la moda de mayor densidad se encuentra en el intervalo $(L_{i-1}\ ,\ L_i]$ que posee la mayor densidad de frecuencia (el rectángulo más alto). Si las dos clases adyacentes: la anterior y la siguiente, tienen igual densidad de frecuencias, se puede suponer que la moda (máximo relativo) se encuentra en el punto medio de la clase que contiene la moda; en caso contrario la moda estará desplazada un poco hacia la clase adyacente de mayor densidad de frecuencia. (suena razonable este criterio).

Por esta razón se conviene que la moda corresponde a la proyección del punto *0*, ver la figura 2.12, observe que con este procedimiento la moda estará siempre más cerca de la clase adyacente con mayor densidad de frecuencia.

Con la notación que aparece en el gráfico y sabiendo que los triángulos *AOB* y *DOE* son semejantes, se puede escribir:

$$M_0 = L_{i-1} + r$$

Además

$$\frac{\Delta_1}{\Delta_2} = \frac{OG}{OF} = \frac{r}{C_i - r} \quad , \quad \text{de donde :}$$

$$r = \frac{\Delta_1}{\Delta_1 + \Delta_2} * C_i$$

De esta manera

$$M_0 = L_{i-1} + \frac{\Delta_1}{\Delta_1 + \Delta_2} * C_i \qquad (*)$$

Como puede apreciarse del gráfico Δ_1 y Δ_2 corresponden a las diferencias de densidad de frecuencia de la clase $(L_{i-1}, L_i]$ con la anterior y con la siguiente respectivamente, ésto es:

$$\Delta_1 = \frac{f_i}{C_i} - \frac{f_{i-1}}{C_{i-1}}$$

$$\Delta_2 = \frac{f_i}{C_i} - \frac{f_{i+1}}{C_{i+1}}$$

Reemplazando Δ_1 y Δ_2 en la expresión (*) tenemos:

$$M_0 = L_{i-1} + \frac{\frac{f_i}{C_i} - \frac{f_{i-1}}{C_{i-1}}}{\frac{2f_i}{C_i} - \frac{f_{i-1}}{C_{i-1}} - \frac{f_{i+1}}{C_{i+1}}} * C_i$$

Donde: $(L_{i-1}, L_1]$: clase que contiene la moda

f_j = frecuencia relativa del intervalo $(L_{j-1}, L_j]$

$C_j = L_j - L_{j-1}$ longitud del intervalo *j-ésimo*

En la siguiente página se presenta un ejemplo del cálculo de la moda.

Ejemplo 2.12

Calcular la moda, a partir del siguiente cuadro de frecuencias:

Intervalo de clase	Frecuencia absoluta (n_i)	Frecuencia relativa (f_i)	Densidad de frecuencia f^*_i
(10,20]	15	7.5	0.75%
(20,40]	40	20	1%
(40,70]	90	45	1.5%
(70,75]	5	2.5	0.5%
(75,100]	50	25	1%
	200	100%	

Como puede apreciarse la clase de mayor densidad de frecuencia es (40, 70] así pues que:

$$M_0 = 40 + \frac{1.5\% - 1\%}{(1.5\% - 1\%) + (1.5\% - 0.5\%)} \times 30$$

$$M_0 = 50$$

La moda se usa con mucha frecuencia como indicador de centralidad en características que tienen escala nominal débil, como la escala nominal u ordinal, no obstante tiene grandes aplicaciones en variables continuas de escala fuerte, por ejemplo en biología, cuando se quiere asociar por ejemplo edad y longitud de peces, seguir el comportamiento de la moda en el tiempo, es una manera de hacer seguimiento a una cohorte de peces. Una aplicación extraordinariamente importante de la moda, la constituye el llamado método de la máxima verosimilitud para

construir estimadores, que es muy usado por sus importantes propiedades asintóticas. El resultado de estos métodos es el hallazgo de la moda de una función de probabilidad o de densidad, llamada función de verosimilitud.

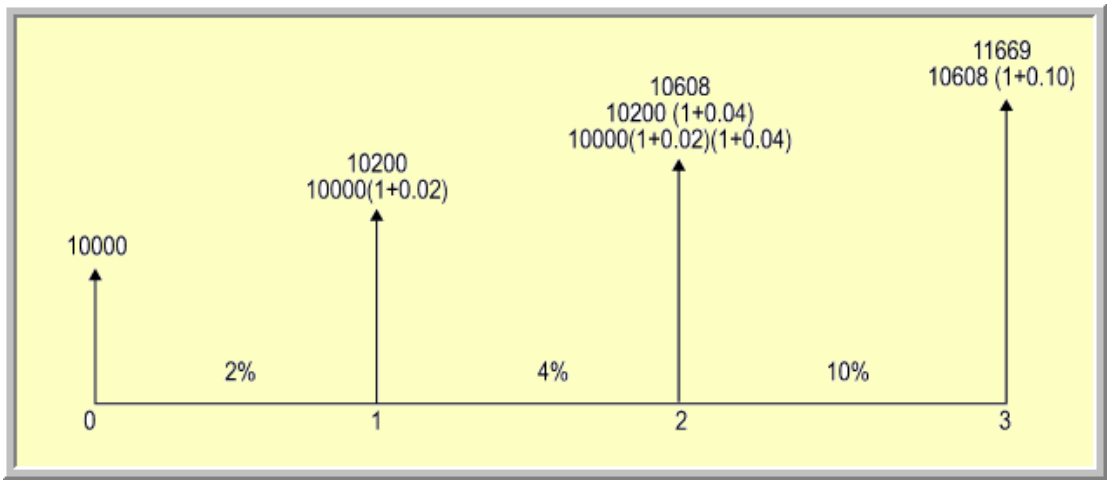
2.4.1.4 La media geométrica

Para tratar de comprender mejor el sentido de la definición de la media geométrica, se presenta el siguiente ejemplo:

Ejemplo 2.13

Una población que tenía 10.000 habitantes en el año cero, creció el primer año a una tasa del 2%, el segundo año creció a una tasa del 4% y el tercer año al 10%. ¿Cuál es el factor de expansión promedio de la población en los 3 años?

Veamos el siguiente esquema:

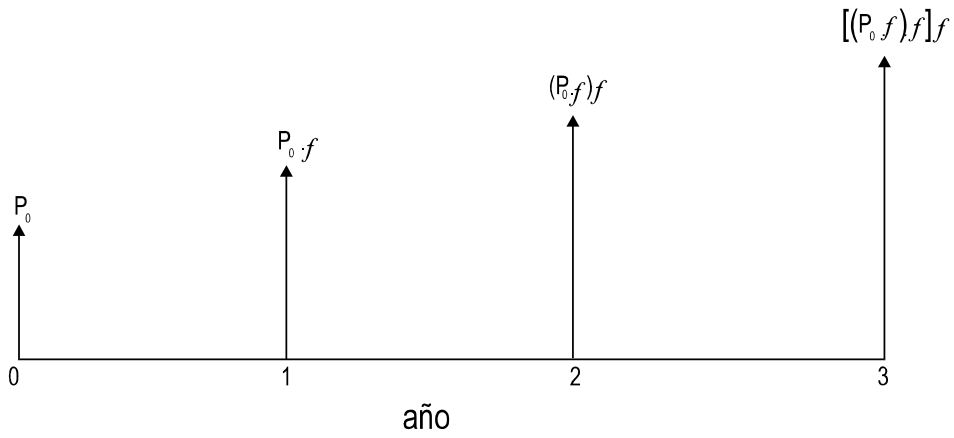


Lo cual significa que al final del año 1, la población era de 10.200 habitantes, es decir se multiplicó por el factor de expansión $f_1 = (1 + 0.02) = 1.02$, al siguiente año, los 10.200 crecieron en un 4% para quedar al final del año 2 una población de 10.608, es decir que los 10.200 se multiplicaron por el factor de expansión $f_2 = (1 + 0.04) = 1.04$; por último los 10.608 se multiplicaron por el factor de expansión $f_3 = (1 + 0.10) = 1.10$ para resultar al final del tercer año, una población de 11.669 habitantes es decir que:

$$P_3 \;=\; P_0 \cdot f_1 \cdot f_2 \cdot f_3 \;=\; 11.669$$

El factor *f* de expansión promedio debe ser tal que comenzando con la misma población P_0 y expandiéndose por el mismo factor *f* todos los años, al final del tercer año debe obtenerse la misma población P_3 que producen los factores $f_1, f_2, y\ f_3$.

Veamos como actuaría *f* promedio, en el siguiente esquema:



Es decir que si la población se expandiera cada año por el mismo factor *f,* la población al final del tercer año será: $P_0 f^3$ que debe ser equivalente con la aplicación de los factores f_1, f_2, f_3, o sea:

$$P_0 \cdot f^3 \;=\; P_0 \cdot f_1 \cdot f_2 \cdot f_3$$

Así que: $$f = \sqrt[3]{f_1 \cdot f_2 \cdot f_3}$$

Decimos aquí que *f* es la **media geométrica,** de $f_1, f_2\ y\ f_3$

Con los números del ejemplo, la media geométrica de los factores de expansión: 1.02, 1.04, 1.10 es:

$$f = \sqrt[3]{1.02 \times 1.04 \times 1.10} = 1.0527$$

Lo cual implica que la tasa de crecimiento promedia de la población fue 5.27%

Generalizando se dirá que la media geométrica M.G. de los datos $x_1, x_2, ...,x_n$ es:

$$\text{M.G.} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Si la variable x es discreta y se conoce su distribución de frecuencias, entonces puede escribirse como:

$$\text{M.G.} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_m^{n_m}}$$

Y si los datos están agrupados en intervalos de clase puede escribirse como:

$$\text{M.G.} = \sqrt[n]{\left(x_1^{'}\right)^{n_1} \cdot \left(x_2^{'}\right)^{n_2} \cdot \ldots \cdot \left(x_m^{'}\right)^{n_m}}$$

2.4.2 Indicadores de dispersión

En la sección anterior se consideraron algunos indicadores de tendencia central, que se pretende fueran representantes de la magnitud de los datos de la muestra; pero el nivel de representatividad de estas medidas, depende del grado de homogeneidad o de dispersión de los datos en la muestra, por tanto se hace necesario estudiar algunos indicadores de dispersión, con el objeto de tener una medida de confianza en los indicadores de centralidad; considere las siguiente situación:

Se tiene dos grupos de datos, el grupo *A*: 2, 98, 3, 97, y el grupo *B*: 49, 51, 48, 52; obsérvese que aunque en ambos grupos el promedio es 50, da la impresión de que este promedio representa mejor los datos del grupo *B* que los del grupo *A*, puesto que los datos del grupo *B* están menos dispersos.

Algunas de las medidas de dispersión más importantes son las siguientes:

2.4.2.1 El rango. (*r*)

Está definido por la distancia entre el menor y el mayor de los datos:

$$r = \max(x_i) - \min(x_i)$$

Por ejemplo, en la muestra: 2, 4, 3, 1, 7, 1, 11, 2, 3, 94. El rango es $r = 94 - 1 = 93$

El rango es sencillo de calcular y de muy fácil interpretación, pero tiene la gran desventaja que es demasiado sensible a valores extremos, en el ejemplo se observa que todos los datos, excepto el 94, están entre 1 y 11, sin embargo, un valor extremo (94) hace que el rango sea 93.

2.4.2.2 La desviación media (D.M)

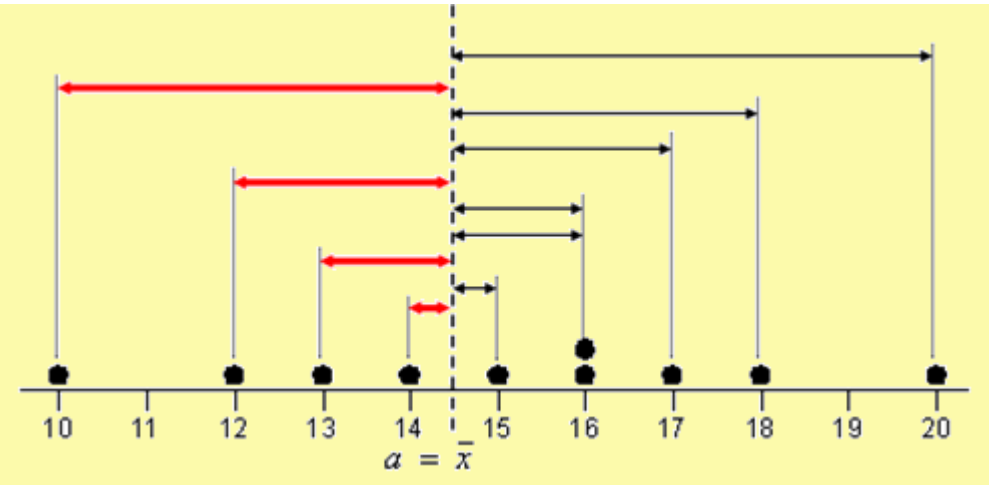
Es un indicador de dispersión que corresponde a la distancia promedio de los datos a la mediana.

$$\text{D.M.} = \frac{\sum_{i=1}^{n} |x_i - \text{Me}|}{n}$$

Si se dispone de una distribución de frecuencias, donde cada x_i aparece asociado con su frecuencia n_i, entonces puede escribirse:

$$\text{D.M.} = \frac{\sum_{i=1}^{m} n_i |x_i - \bar{x}|}{n} = \sum_{i=1}^{m} f_i |x_i - \bar{x}|$$

Que corresponde a la media de las distancias que se presentan en el gráfico que esta a continuación:



Si los datos están agrupados en intervalos de clase, una expresión aproximada para el cálculo de la desviación media es:

$$\text{D.M.} = \sum_{i=1}^{m} \frac{n_i \left| x_i' - \bar{x} \right|}{n} = \sum_{i=1}^{m} f_i \left| x_i' - \bar{x} \right|$$

Donde x_i' es la marca de clase de intervalo *i.*

En la muestra: 2, 5, 8, 1, 4 cuya mediana es Me = *4*, la desviación media es:

$$\text{D.M.} = \frac{|2-4| + |5-4| + |8-4| + |1-4| + |4-4|}{5} = 2$$

Lo cual indica que en promedio los datos están separados de la mediana M_e en 2 unidades.

La desviación media es un indicador de fácil interpretación directa, pero su estructura matemática (valor absoluto) ha hecho difícil su uso en los desarrollos inferenciales de la estadística, en cambio existen otros que superan esta dificultad y por tal razón están asociados con muchos procedimientos de la inferencia, como por ejemplo:

2.4.2.3 La varianza (S^2)

Esta es la medida de dispersión más usada en estadística y está definida como:

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

Si se dispone de una distribución de frecuencias $\{(x_i, n_i)\}$, se pueden calcular como:

$$S^2 = \frac{1}{n} \sum_{i=1}^{m} n_i \left(x_i - \bar{x} \right)^2 = \sum_{i=1}^{m} f_i \left(x_i - \bar{x} \right)^2$$

Si los datos están agrupados en intervalos de clase, una expresión aproximada para la varianza es:

$$S^2 = \frac{1}{n} \sum_{i=1}^{m} n_i \left(x_i' - \bar{x} \right)^2 = \sum_{i=1}^{m} f_i \left(x_i' - \bar{x} \right)^2$$

No obstante que la varianza está dada por una expresión cuadrática, que ofrece muchas ventajas en la manipulación matemática, tiene algunas desventajas, entre las cuales están: su no fácil interpretación directa y que sus unidades no coinciden con las unidades de la variable en estudio, así por ejemplo si x está en metros, su varianza estará dada en metros cuadrados. Esta última desventaja se pretende remediar extrayendo la raíz cuadrada a la varianza para obtener la que se conoce como **desviación estándar** (S), que será:

$$S = \sqrt{\frac{1}{n}\sum\left(x_i - \bar{x}\right)^2}$$

Interpretación de la desviación estándar (principio de Tchebychev)

Una interpretación de la desviación estándar puede hacerse a través del **principio de Tchebychev**) que expresa que para cualquier muestra $x_1, x_2, ...,x_n$ se cumple que si se construye un intervalo con centro en la media y con extremos ubicados a una distancia de k veces la desviación estándar S, en dicho intervalo está por lo menos $(1 - \frac{1}{k^2})$ x 100% de los datos; escrito en símbolos será:

$$f\left(\bar{x} - ks, \bar{x} + ks\right) \geq 1 - \frac{1}{k^2}$$

Así por ejemplo si $k = 2$, dice que:

$$f\left(\bar{x} - 2s, \bar{x} + 2s\right) \geq 1 - \frac{1}{2^2} = 0.75$$

Es decir que en el intervalo construido a 2 desviaciones estándar a cada lado de la media está por lo menos el 75% de los datos. Para $k = 3$, se dice que está por lo menos el 88.8% de los datos.

Este principio proporciona cotas para la frecuencia, en términos de la desviación estándar, lo cual ayuda a su interpretación, pero como es muy general, dichas cotas pueden ser muy bajas, se

observa que para $k = 1$ el principio dice que en el intervalo $(\bar{x} - s\,,\, \bar{x} + s)$ hay por los menos el 0% de los datos, lo cual es obvio.

Propiedades de la varianza

Las propiedades que se presentan a continuación pueden ser heredadas por la desviación estándar con las limitaciones que genera la función raíz cuadrada.

1. $S^2 = \dfrac{\sum_i^n x_i^2}{n} - \left(\bar{x}\right)^2$

Esta, más que una propiedad es una forma alternativa de calcular la varianza, realizando menos cálculos numéricos que con la expresión que proporciona la definición. Su demostración es la siguiente:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[x_i^2 - 2\bar{x}x_i + \left(\bar{x}\right)^2\right]$$

$$= \frac{1}{n}\sum x_i^2 - \frac{1}{n}\cdot 2\bar{x}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\left(\bar{x}\right)^2$$

$$= \frac{1}{n}\sum x_i^2 - 2\bar{x}\cdot\frac{\sum_{i=1}^{n} x_i}{n} + \frac{1}{n}\cdot n\left(\bar{x}\right)^2$$

$$= \frac{1}{n}\sum x_i^2 - 2\left(\bar{x}\right)^2 + \left(\bar{x}\right)^2$$

$$S^2 = \frac{1}{n}\sum x_i^2 - \left(\bar{x}\right)^2$$

S^2 = Promedio de los cuadrados, menos, promedio al cuadrado

2. La varianza es siempre no negativa.

$S^2 \geq 0$, esto se desprende de que la varianza es una suma de cuadrados, multiplicada por la constante, $\frac{1}{n}$, que siempre es positiva.

3. La varianza de una constante es cero, es decir: si $x_i = C$, para todo i, entonces

$S_x^2 = 0$

$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, pero se sabe que si $x_i = C$ entonces $\bar{x} = C$, de este modo:

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(C - C\right)^2 = 0$$

4. Si $y_i = kx_i$, entonces $S_y^2 = k^2 S_x^2$ $\quad i = 1, 2, ..., n$

Es decir: si se tiene una muestra $x_1, x_2, ..., x_n$, que tiene varianza S^2_x y cada dato se multiplica por la constante k, la varianza de esta nueva muestra:

$$Kx_1, Kx_2, ..., Kx_n, \text{ será } k^2 S_x^2$$

lo cual puede demostrarse de la siguiente manera:

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(kx_i - k\bar{x}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} K^2\left(x_i - \bar{x}\right)^2 = K^2 \cdot \frac{1}{n}\sum\left(x_i - \bar{x}\right)^2$$

$$= K^2 S_x^2$$

5. Si $y_i = x_i + C$, entonces $S_y^2 = S_x^2$

$i = 1, 2, ..., n$

Es decir: que si todos los datos se trasladan la misma distancia C, la varianza no cambia, lo cual puede verificarse así:

$$S_y^2 = \frac{1}{n}\sum\left(y_i - \bar{y}\right)^2 = \frac{1}{n}\sum\left[\left(x_i + C\right) - \left(\bar{x} + C\right)\right]^2$$

$$= \frac{1}{n}\sum\left(x_i + C - \bar{x} - C\right)^2 = \frac{1}{n}\sum\left(x_i - \bar{x}\right)^2$$

$$= S_x^2$$

Ejemplo 2.14

Dada la siguiente distribución de frecuencias sobre una variable continua x, que se presenta en el cuadro, en el que se registra: el intervalo de clase (X_i'), las frecuencias absolutas y las frecuencias relativas.

Intervalos de clase	Marca de clase (x'_i)	Frecuencia absoluta (n_i)	Frecuencia relativa (f_i)
(10,20]	15	20	0.10
(20,40]	30	70	0.35
(40,70]	55	50	0.25
(70,80]	75	40	0.20
(80,100]	90	20	0.10
		200	1.00

a) Calcule la desviación media

$$\text{D.M.} = \frac{\sum_{i=1}^{m} n_i \left|x_i' - M_e\right|}{n}$$

Se debe calcular primero M_e

$$M_e = L_{i-1} + \frac{0.50 - H(L_{i-1})}{f_i} * C_i$$

$$M_e = 40 + \frac{0.50 - 0.45}{0.25} * 30 = 46$$

Entonces

$$\text{D.M.} = \frac{20|15 - 46| + 70|30 - 46| + 50|55 - 46| + 40|75 - 46| + 20|90 - 46|}{200}$$

D.M. $\approx$ 21.15

Lo cual indica que en promedio los datos están separados de la mediana en 21.15 unidades

b) Calcule la varianza

Usando la forma simplificada:

S^2 = Promedio de los cuadrados, menos, promedio al cuadrado

$$= \frac{\sum^{m} n_i (x_i')^2}{n} - (\bar{x})^2$$

$$= \frac{20 \times (15)^2 + 70 \times (30)^2 + \ldots + 20 \times (90)^2}{200} - (49.75)^2$$

$$= 3.028{,}7 - 2.475{,}1 = 553{,}7$$

c) La desviación estándar

$$S = \sqrt{553.7} = 23.5$$

d) Verifique el principio de Tchebychev para $k = 2$, es decir se debe verificar que:

$$f\left(\bar{x} - 2s, \bar{x} + 2s\right) \geq 1 - \frac{1}{2^2} = 0.75 \equiv 75\%$$

$$\bar{x} - 2s = 49.75 - 2(23.5) = 2.75$$

$$\bar{x} + 2s = 49.75 + 2(23.5) = 96.75$$

Estimando de acuerdo con la tabla de frecuencias qué porcentaje de datos hay en el intervalo (2.75 , 96.75):

$$f(2.75\ ,\ 96.75) = 0.10 + 0.35 + 0.25 + 0.20 + \frac{0.10}{20} \times 16.75$$

$$= 0.984 > 0.75$$

2.4.2.4 El coeficiente de variación

Por la estructura de la varianza se sabe que cuando aumenta la dispersión el valor de la varianza aumenta, por esa razón se usa como indicador de dispersión, igualmente la desviación estándar; pero, qué se respondería a la pregunta: ¿una desviación estándar de 200 metros es grande o es pequeña ? o de otra manera: ¿una desviación estándar de 200 metros me indica que hay poca o mucha dispersión ?

La respuesta casi obligada es: depende..., porque si las magnitudes de los datos de la respuesta son "grandes", por ejemplo: la distancia recorrida diariamente por un cartero, registrada durante 30 días. En este caso, una desviación estándar de 200 metros puede ser pequeña, así como una desviación estándar de 10 micras podría ser grande si se está estudiando el diámetro de ciertas células.

Lo anterior muestra la necesidad de definir un indicador de dispersión que involucre la magnitud de los datos que se estudian; magnitud ésta que puede ser representada por la media aritmética, esto da origen al llamado: c**oeficiente de variación**, que consiste en expresar la desviación estándar como un porcentaje de la media aritmética, así pues:

$$\text{C.V.} = \frac{S}{\bar{x}} \times 100\%$$

Entonces, si una muestra tiene una media aritmética $\bar{x}$ = 40.000 metros y una desviación estándar $S = 500$ metros entonces:

$$\text{C.V.} = \frac{500}{40.000} \times 100\% = 1.25\%$$

Que podría indicar una dispersión relativamente pequeña.

En realidad el coeficiente de variación se usa para comparar la variabilidad relativa de una característica, en poblaciones que tienen distinta media.

No existen topes, que permitan valorar un coeficiente de variación como grande o pequeña.

El juicio sobre su tamaño esta siempre ligado al problema específico que se estudia. Surgen de esta manera y como resultado de la propia experiencia en un campo específico, valores de coeficiente de variación como limitantes en un proceso de control de calidad. Algunas de las normas sobre materiales de construcción exigen no sólo un promedio de resistencia por encima de un nivel mínimo, sino también control sobre la variabilidad expresado en forma de coeficiente de variación.

En el área de la agricultura, una determinada variedad de maíz puede tener asociado (por la experiencia) como natural, un coeficiente de variación en su rendimiento por hectárea, cual es distinto (generalmente menor) si el cultivo está bajo riego, que si esta bajo temporal.

El coeficiente de variación, puede ser característica de un fenómeno en especial. Se sabe por ejemplo que si la función de densidad de frecuencia de una característica tiene forma exponencial, siempre su coeficiente de variación es de 100%, como consecuencia de que la media y la desviación estándar son iguales en esta familia de distribuciones.

EJERCICIOS PROPUESTOS

1. Una entidad encargada del control de contaminación de cierto río, lleva registros sobre el oxígeno disuelto, *X*, expresado en *mg/l*; éstos se presentan a continuación:

   2.6, 3.6, 3.1, 2.6, 2.7, 3.9, 2.4, 2.7, 2.5, 2.3, 4.0, 3.2, 2.5, 1.7, 0.3, 3.1, 2.6, 1.3, 4.3, 1.5, 2.8, 1.8, 4.2, 3.5, 2.4, 2.2, 3.4, 3.7, 0.8, 2.3, 1.9, 4.5, 1.2, 2.2, 2.2, 3.0, 2.1, 1.8, 2.9, 3.8, 3.5, 1.6, 3.2, 4.4, 1.4, 0.7, 2.8, 3.3, 0.5, 2.3 .

   a) Agrupe la información en intervalos de clase y construya un cuadro de frecuencias completo.

   b) Grafique el histograma, y la ojiva.

   c) Calcule el porcentaje de registros que son inferiores a 3.1 *mg/l.*

      c.1 Usando la ojiva

      c.2 A partir del cuadro de frecuencias

      c.3 Por conteo directo de la muestra bruta

      Compare los resultados y comente.

   d) Estime el porcentaje de registro que son mayores que 1.5 *mg/l*, pero son menores que 3.5 *mg/l.*

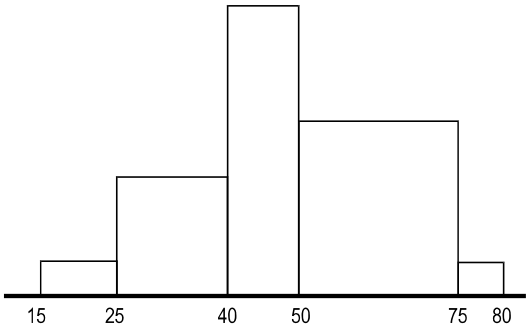
e) Calcule la media aritmética, la mediana y la moda.

f) Calcule la desviación estandar. ¿Le parece grande? Justifique.

g) Qué porcentaje de los registros están entre $\bar{x} - 2S$ y $\bar{x} + 2S$? Se cumple aquí el principio de Tchebycheff ?

h) Construya un diagrama de cajas y alambres e interprete.

2. Dada la información que proporciona el siguiente gráfico, estime el porcentaje de datos que son mayores de 27 pero menores que 52.



3. Si en una muestra de 50 datos, se obtuvo: $\bar{x} = 50$ y $S^2 = 100$ y se recogieron a última hora los siguientes datos adicionales: 32, 84, 36, 51, 23, ¿cuál es la nueva media y la nueva varianza?

4. Verifique si:

$$Z_i = \frac{X_i - \overline{X}}{S_x} \quad ; \quad i = 1, 2, \ldots, n$$

Entonces: $\overline{Z} = 0$ y $S_z^2 = 1$

5. Decida sobre la VERACIDAD o FALSEDAD de las siguientes proposiciones, justificando claramente la razón de su decisión:

   a) Si las frecuencias absolutas de los datos de una muestra se triplican su media aritmética no variaría.

   b) Si a cada uno de los datos de una muestra se le sumara 3 unidades y su frecuencia absoluta se triplicara su desviación estándar no cambiaría.

   c) Si a cada uno de los intervalos de clase de una tabla de frecuencias se le agregan tres datos, la mediana podría cambiar pero la moda no.

   d) La media aritmética de la muestra bruta debe coincidir siempre con la media aritmética calculada con base en los datos agrupados.

   e) Si una muestra se divide en 2 subgrupos n_1 y n_2 elementos $(n_1 + n_2 = n)$, con varianzas S_1^2 y S_2^2 respectivamente, entonces la varianza de la muestra puede expresarse como:

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$$

   f) Si a los datos: $x_1, x_2, ..., x_n$, de una muestra se aplica la transformación

   $y_i = ax_i + b$, con $a > 0$ y $b > 0$, entonces $"y"$ tiene menor dispersión relativa que $"x"$ (en términos del coeficiente de variación).

6. Si $P_1, P_2, ...,P_n$ representa la población (número de habitantes) de una región en los años 1, 2, ...,n respectivamente usando el concepto de media geométrica, encuentre una expresión para estimar la tasa de crecimiento. Obsérvela y comente las ventajas que presenta.

7. En una población del Cauca se tomó una muestra de 50 familias para observar el número de personas menores de 12 años con el propósito de estimar algunos indicadores sobre demanda potencial de educación escolar. Esta arrojó los siguientes resultados:

4	0	1	2	3	0	2	5	3	1
3	2	1	2	1	3	0	3	0	1
0	2	3	0	1	4	2	1	5	4
2	1	4	2	3	1	2	0	1	3
2	2	5	0	3	3	2	0	1	5

7.1 Con base en la información anterior llene la siguiente tabla de frecuencias.

X_i	Frecuencia absoluta (n_i)	Frecuencia relativa (f_i)	Frecuencia acumulada	
			absolutas	relativas

7.2 Determine qué porcentaje de las familias tienen 3 personas o menos que son menores de 12 años.

7.3 Si la población consta de 1.200 familias estime usted, el número de personas menores de 12 años.

7.4 Usted está seguro del resultado obtenido en el numeral 7.3 ? qué supuesto está implícito en la estimación?

7.5 Construya un gráfico para la distribución empírica de frecuencias acumuladas relativas.

8. Una compañía constructora resuelve estudiar en un concreto su resistencia a la compresión, con el objeto de hacer un control de calidad. Para ello se tomaron 50 cilindros de prueba de acuerdo con las normas establecidas. Los resultados en kg/cm^2 de presión obtenidos al cabo de 28 días de curado fueron:

295, 282, 221, 240, 328, 191, 275, 253, 259, 206, 265, 253, 237, 290, 251, 199, 254, 217, 247, 296, 240, 230, 312, 226, 335, 232, 220, 214, 272, 245, 212, 241, 234,226, 309, 226, 221, 248, 264, 268, 250, 268, 289, 277, 283, 273, 214, 269, 253, 252.

8.1 Llene la siguiente tabla de frecuencias:

Intervalo de clase	Marca de clase	C_i	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	
					absoluta	relativa

8.2 Especifique la función empírica de densidad de frecuencias

8.3 Especifique la función empírica de distribución acumulada relativa.

8.4 Calcule el porcentaje de cilindros que resistieron más de 235 kg/cm^2 pero menos 264 kg/cm^2.

8.5 Estime el riesgo, si se usa ese concreto en una obra que exige 240 kg/cm^2 de resistencia a la compresión. Le parece alto ?

8.6 Calcule con base en los datos agrupados:

a) La media aritmética

b) La mediana

c) La moda

8.7 Calcule con base en los datos agrupados la desviación estándar.

8.8 Le parece grande la dispersión? Justifique.

8.9 Si se entera que el equipo de medición de resistencia tiene un error sistemático, en el sentido que muestra una lectura superior en 5 kg/cm^2 al verdadero valor, entonces calcule la media aritmética, la mediana, la moda y la desviación estándar reales, a partir de los puntos 8.6 y 8.7.

8.10 Si el error sistemático consistiera en amplificar el valor real en un 10%. Calcule la media y desviación estándar reales.

8.11 Si se aumenta la muestra con 10 cilindros más que se prueban con los siguientes resultados: 232, 256, 287, 228, 295, 226, 277, 233, 247, 277.

Calcule la nueva media y la nueva varianza, usando los resultados encontrados en 8.6 y 8.7.

8.12 Construya un diagrama de cajas y alambres para los datos originales e

Interprete.

9. Si la característica X de una población tiene la siguiente función de densidad:

$$f^*(x) = \begin{cases} ax & \text{si } 0.0 \leq x < 0.5 \\ a & \text{si } 0.5 \leq x \leq 1.5 \\ 0 & \text{en otra parte} \end{cases}$$

a) Encuentre el valor adecuado para la constante *"a"*.

b) Calcule el porcentaje de datos que cumplen que $0.3 < x \leq 1.1$.

c) Si se tomara una muestra al azar de 10.000 elementos de dicha población, ¿Cuántos de ellos, esperaría usted tengan la característica *X* en el intervalo

(0.3 , 1.1] ?

10. Dada la siguiente información sobre el crecimiento de una población:

Año	Tasa de crecimiento por cada 1000 habitantes
1978	+28
1979	+34
1980	+20
1981	+33
1982	-33

a) Estime la tasa promedia de crecimiento

b) Haga una proyección de la población para 1988 si se sabe que en 1982 había 102.800 habitantes.

c) Estime el número promedio de años que deben transcurrir para que dicha población tenga 500.000 habitantes?

Capítulo 3

DISTRIBUCIONES BIDIMENSIONALES DE FRECUENCIA

3.1 INTRODUCCIÓN

En el capítulo anterior nos ocupamos del tratamiento descriptivo de datos correspondientes a la observación de una característica en los elementos que constituían el objeto de estudio. En ocasiones es de interés hacer el tratamiento conjunto de dos características o variables observadas en los elementos de una muestra o de una población, por ejemplo, puede ser importante considerar en forma simultánea las características: "costos" y "producción" por hectárea cultivada de plátano, en las fincas del Valle del Cauca. En otra situación podría ser útil considerar conjuntamente las variables: "número de personas que habitan" y "área de dormitorio" para las viviendas de la población de Guachené. En el campo industrial por ejemplo: "hora del día" y "número de artículos defectuosos producidos". En el área de la salud: "edad" y "peso" de los niños de cierta comunidad. En Biología: "consumo de alimento" y "ganancia de peso" de los pollos de una granja experimental. En ingeniería: "caudal" y "profundidad" en cierto punto del cauce de un río. En

Economía: "precio unitario" y "demanda" de cierto artículo. En Educación: "estrato socioeconómico" y "rendimiento académico" de los estudiantes de educación primaria en la ciudad de Palmira. En el campo de la Sociología: "ingreso percapita" e "indice de criminalidad" en las poblaciones de Colombia, también podría ser de interés estudiar las variables: "indice de analfabetismo" e "indice de criminalidad". En el área de la salud pública: "tasa de mortalidad infantil" y "cobertura de abastecimiento de agua tratada" en un conjunto de poblaciones de la región occidental de Colombia. Para el médico rural sería útil establecer relaciones entre: "consistencia de las heces fecales" y "presencia de cierto tipo de parásito". En el campo de la administración: "plazo en los créditos" y "mora en los pagos" o también "volumen de ventas" y "monto de la cobranza" para distintos meses del año.

En las situaciones mencionadas, puede interesar al investigador, la distribución de frecuencias, considerando **conjuntamente** los diferentes valores (o categorías) de las variables. Puede ser de interés considerar el comportamiento estadístico de una variable para los elementos que tienen un determinado valor en la otra variable considerada. En ocasiones es útil explorar sobre el grado de asociación de dos características en los elementos de cierta población. También puede requerirse "predecir" el valor de una característica de un elemento en particular, aprovechando el conocimiento de otra característica del mismo elemento, valiéndonos de la asociación estadística que exista entre ellas.
En el desarrollo del presente capítulo vamos a ocuparnos de dar respuesta a esas situaciones.

3.2 DISTRIBUCIONES CONJUNTAS Y DISTRIBUCIONES MARGINALES

En los ejemplos mencionados en la introducción de este capítulo, podemos observar varias situaciones en cuanto a la naturaleza de las variables que se desea estudiar. En algunos casos, ambas características son atributos (variables cualitativas), en otros, ambas son de naturaleza discreta o una de ellas es discreta y la otra continua o ambas son continuas, de acuerdo con la definición que se hizo en el capítulo 2. Esta diferenciación de las variables se hace con el mismo sentido planteado en las distribuciones unidimensionales y será necesario explicitarla sólo en esta primera parte, ya que después, en el tratamiento de otros aspectos en los que no sea determinante su identificación, se dejará implícita la diferencia.

En general se usará la siguiente notación: $X_1, X_2, ..., X_i, ..., X_m$ representan las *"m"* categorías a considerar para clasificar los elementos de la muestra en lo que respecta a la variable X. Estas categorías pueden corresponder a nombres si se trata de escala nominal de las variables cualitativas, puede coincidir con los valores que toma la variable X si es discreta o pueden representar intervalos de clase si X es una variable continua.

Análogamente Y_1, Y_2, ... Y_j, ..,Y_s, representan las *"s"* categorías a considerar para clasificar los elementos de la muestra con respecto a la variable *Y*.

Cuando los elementos de una muestra se clasifican simultáneamente por dos (2) características *X* e *Y*, surge para su representación las llamadas "tablas de doble entrada" que se construirán más adelante.

Se entiende que un elemento de la muestra se clasifica en sólo una categoría de *X* y en sólo una categoría de *Y*.

Si se llama Ω al conjunto de todos los elementos de la muestra y se llama X_i al conjunto de los elementos de la muestra que pertenecen a la *i- ésima* categoría de *X* y análogamente para *Y* entonces:

- $X_i \cap X_k = \emptyset \;\; si\; i \neq k$
- $X_1 \cup X_2 \cup ... \cup X_m = \Omega$
- $Y_j \cap Y_t = \emptyset \;\; si\; j \neq t$
- $Y_1 \cup Y_2 \cup ... \cup Y_s = \Omega$
- $(X_i \cap Y_1) \cup (X_i \cap Y_2) \cup ... \cup (X_i \cap Y_s) = X_i$
- $(X_1 \cap Y_j) \cup (X_2 \cap Y_j) \cup ... \cup (X_m \cap Y_j) = Y_j$
- $\bigcup_{i=1}^{m} \bigcup_{j=1}^{s} \left(X_i \cap Y_j \right) = \Omega$

A continuación se trata en forma particular las distintas situaciones que se presentan, dependiendo de si *X* e *Y* son variables discretas o continuas.

3.2.1 Caso en que ambas variables son de naturaleza discreta

Para ilustrar este caso se plantea el siguiente ejemplo:

Ejemplo 3.1

De cierta población en estudio se sacó una muestra de 50 familias con el propósito de observar las variables: "número de personas que componen la familia" (*X*) y "número de personas que producen algún ingreso" (*Y*), los datos obtenidos presentados como parejas (*X,Y*) son los siguientes:

(6,1), (1,1), (3,1), (4,2), (6,1), (1,1), (3,1), (4,2), (5,2), (5,1), (5,4), (6,1), (2,1), (3,2), (4,3), (6,2), (2,1), (3,2), (4,2), (3,2), (4,2), (4,3), (3,3), (4,3), (4,4), (4,4), (4,4), (4,2), (2,1), (6,2), (6,3), (4,4), (2,1), (5,1), (5,5), (4,4), (3,2), (2,2), (6,4), (6,5), (6,4), (6,2), (6,3), (6,2), (6,2), (5,2), (5,4), (5,1), (5,4), (5,4)

Los datos anteriores pueden ser organizados haciendo conteos en forma análoga a como se hizo en el caso unidimensionales como se muestra a continuación:

CUADRO 3.1

DISTRIBUCION CONJUNTA DE FRECUENCIAS ABSOLUTAS DE LAS VARIABLES "NUMERO DE PERSONAS QUE COMPONEN LA FAMILIA" (*X*) Y "NUMERO DE PERSONAS QUE PRODUCEN ALGUN INGRESO" (*Y*)

X \ Y	Y_1	Y_2	...	Y_j	...	Y_s	
X_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1\cdot}$
X_2	n_{21}	n_{21}	...	n_{2j}	...	n_{2s}	$n_{2\cdot}$
:	:	:	:	:	:	:	:
X_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i\cdot}$
:	:	:	:	:	:	:	:
X_m	n_{m1}	n_{m2}	...	n_{mj}	...	n_{ms}	$n_{m\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot j}$	...	$n_{\cdot s}$	n

X \ Y	1	2	3	4	5	
1	2	0	0	0	0	2
2	4	1	0	0	0	5
3	2	4	1	0	0	7
4	0	5*	3	5	0	13
5	3	2	0	4	1	10*
6	3	5	2	2	1	13
	14	17*	6	11	2	50

El dato (6,3) indica que la familia observada está compuesta por 6 personas de las cuales 3 producen algún tipo de ingreso.

Con respecto a los valores que figuran en el cuadro 3.1, con * pueden interpretarse de la siguiente manera:

- El 5 indica que en la muestra observada se presentaron cinco (5) familias compuestas por 4 personas de las cuales 2 producen algún tipo de ingreso, es decir para las cuales $X = 4$ y $Y = 2$. Entonces $n_{42} = 5$

- El 10 significa que en la muestra hay diez (10) familias compuestas por 5 personas; es decir para las cuales $X = 5$. Entonces $n_{5.} = 10$

- El 17 indica que en la muestra se encontró diecisiete (17) familias en las cuales hay 2 personas que trabajan, es decir para las cuales $Y = 2$. Entonces $n_{.2} = 17$.

A continuación se presenta la representación gráfica de la distribución conjunta del Ejemplo 3.1

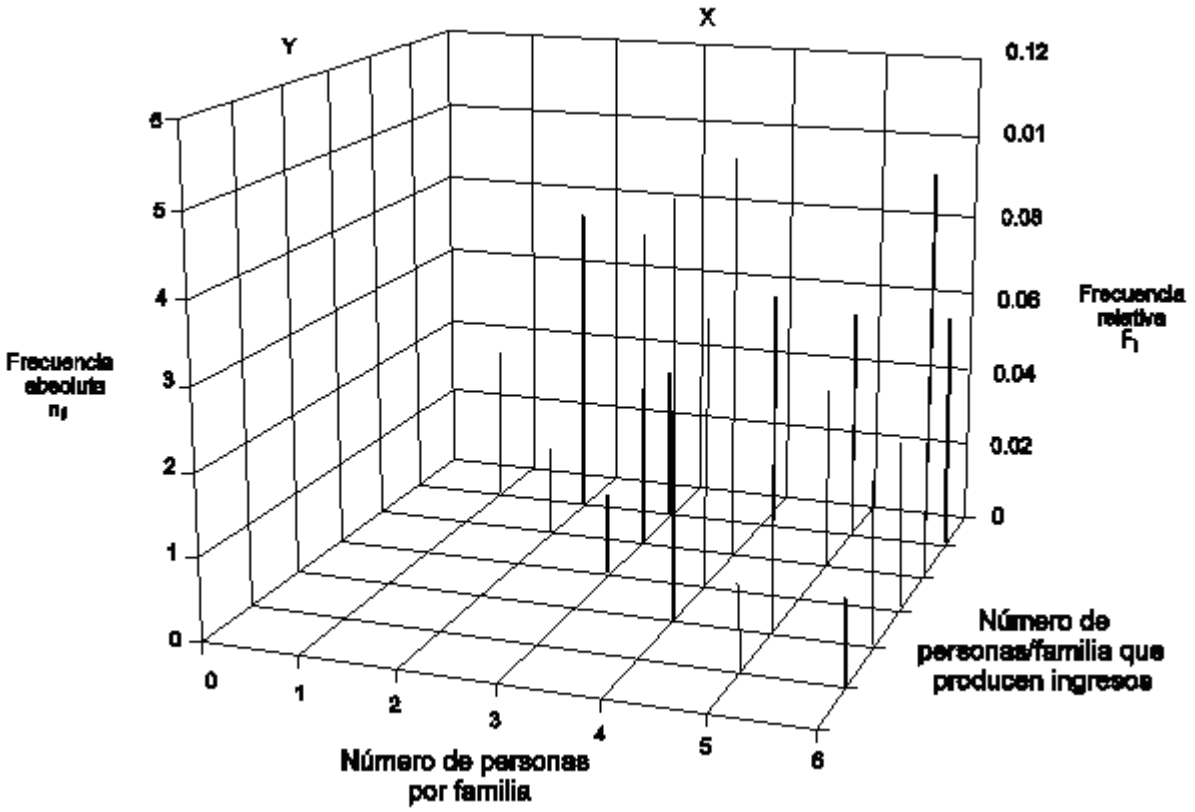


Fig. 3.1: Distribución conjunta de frecuencias absolutas y relativas de las variables "número de personas/familia" (*X*) y "número de personas que producen algún ingreso en la familia" (*Y*).

Si se consideran las frecuencias que aparecen al margen en el cuadro 3.1, se obtiene información sobre una sola variable, bien sea sobre *X* o sobre *Y*, estas distribuciones se les conoce como **distribuciones marginales**.

CUADRO 3.2

DISTRIBUCION MARGINAL DE FRECUENCIAS ABSOLUTAS DE LA VARIABLE "NUMERO DE PERSONAS QUE INTEGRAN LA FAMILIA" (*X*)

Roberto Behar y Mario Yepes

X_i	No. de Familias (Frec. Absoluta)
1	2
2	5
3	7
4	13
5	10
6	13
	50

Notación

X_i	$n_{i.}$
X_1	$n_{1.}$
X_2	$n_{2.}$
:	:
X_i	$n_{i.}$
:	:
X_m	$n_{m.}$
	n

El cuadro 3.2 muestra la distribución de frecuencias de las familias, si sólo se tiene en cuenta el número de personas que las integran.

Análogamente puede construirse la distribución marginal de frecuencias absolutas para la variable "número de personas que trabajan en la familia" (*Y*).

Puede construirse con base en el cuadro 3.1 la distribución conjunta de frecuencias relativas, expresando los números que resulten del conteo, como una fracción o porcentaje del número total de familias observadas (50). Así por ejemplo, el 5 que aparece en el cuadro 3.1 representa el 10% de las 50 familias, así pues la frecuencia relativa asociada al dato (4.2) es 0.10, de esta manera se construye el cuadro 3.3.

CUADRO 3.3

DISTRIBUCION CONJUNTA DE FRECUENCIAS RELATIVAS DE LAS VARIABLES "NUMERO DE PERSONAS QUE COMPONEN LA FAMILIA" (*X*) Y "NUMERO DE PERSONAS QUE PRODUCEN ALGUN INGRESO"(*Y*).

Y / X	1	2	3	4	5	
1	0.04	0.00	0.00	0.00	0.00	0.04
2	0.08	0.02	0.00	0.00	0.00	0.10
3	0.04	0.08*	0.02	0.00	0.00	0.14*
4	0.00	0.10	0.06	0.10	0.00	0.26
5	0.06	0.04	0.00	0.08	0.02	0.20
6	0.06	0.10	0.04	0.04	0.02	0.26
	0.28	0.34*	0.12	0.22	0.04	1.00

En forma general se representa la distribución conjunta de frecuencias relativas de la siguiente manera:

X \ Y	Y_1	Y_2	...	Y_j	...	Y_s	
X_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1s}	$f_{1.}$
X_2	f_{21}	f_{21}	...	f_{2j}	...	f_{2s}	$f_{2.}$
:	:	:	:	:	:	:	:
X_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{is}	$f_{i.}$
v:	:	:	:	:	:	:	:
X_m	f_{m1}	f_{m2}	...	f_{mj}	...	f_{ms}	$f_{m.}$
	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.s}$	1.00

donde $f_{ij} = \frac{n_{ij}}{n}; f_{i.} = \frac{n_{i.}}{n}$

La interpretación de los valores que se destacan en el cuadro 3.3 es la siguiente:

- 0.08 indica el 8% de las familias están compuestos por 3 personas y 2 personas producen algún ingreso, es decir que el dato (3,2) representa el 8% de las 50 observaciones realizadas. Entonces $f_{32} = 0.08$

- 0.14 indica que el 14% de las familias están compuestas por 3 personas; es decir que para el 14% de las familias se cumple que $X = 3$. Entonces $f_{3.} = 0.14$

- 0.34 indica que en el 34% de las familias ocurre que 2 personas producen algún ingreso; es decir que para el 34% de las familias se cumple que $Y = 2$. Entonces $f_{.2} = 0,34$

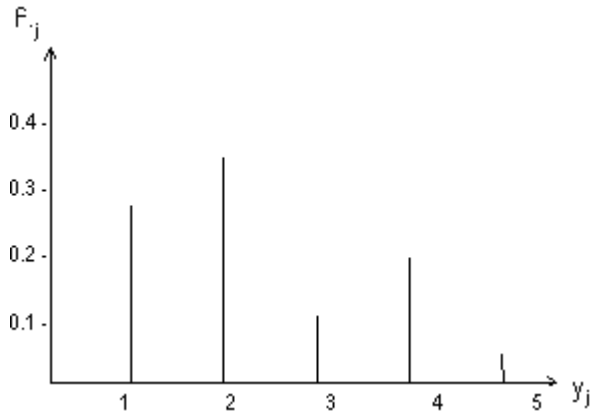


Fig. 3.2: Distribución marginal de frecuencia relativa de la variable "número de personas / familia, que producen algún ingreso" (*Y*).

De nuevo aquí si se considera las frecuencias relativas que aparecen al margen en el cuadro 3.3, se obtiene la llamada **distribución marginal de frecuencias relativas.**

CUADRO 3.4

DISTRIBUCION MARGINAL DE FRECUENCIAS RELATIVAS DE LA VARIABLE "NUMERO DE PERSONAS QUE PRODUCEN INGRESO EN LA FAMILIA" (*Y*).

Y_i	Fracción del Total de Familias
1	0.28
2	0.34
3	0.12
4	0.22
5	0.04
	1.00

Y_j	$f_{\bullet j}$
Y_1	$f_{\bullet 1}$
Y_2	$f_{\bullet 2}$
:	:
Y_j	$f_{\bullet i}$
:	:
Y_s	$f_{\bullet s}$
	1.00

El cuadro 3.4, muestra la distribución relativa de las familias si sólo se observa "el número de personas que producen algún ingreso a la familia".

Puede determinarse el número de familias que tienen 4 miembros o menos y de los cuales trabajan 2 personas o menos, en el ejemplo cumplen con esto, 18 familias, que representan el 36% del número total de familias muestreadas.

Esta situación introduce el concepto de **distribución conjunta de frecuencias acumuladas,** que puede denotarse como *N(x,y)* o como *F(x,y)* según se trate de frecuencias absolutas o relativas acumuladas, como se muestra en el cuadro 3.5.

CUADRO 3.5

DISTRIBUCION CONJUNTA DE FRECUENCIAS RELATIVAS ACUMULADAS DE LAS VARIABLES "NUMERO DE PERSONAS QUE COMPONEN LA FAMILIA" (*X*) Y "NUMERO DE PERSONAS QUE PRODUCEN ALGUN INGRESO A LA FAMILIA (*Y*).

X \ Y	1	2	3	4	5
1	0.04	0.04	0.04	0.04	0.04
2	0.12	0.14	0.14	0.14	0.14
3	0.16	0.26	0.28	0.28	0.28
4	0.16	0.36	0.44	0.54	0.54
5	0.22	0.46	0.54*	0.72	0.74
6	0.28	0.62	0.74	0.96	1.00

El valor 0.54 marcado en el cuadro 3.5 indica que el 54% de las familias están compuestas por 5 ó menos personas de las cuales 3 ó menos aportan algún ingreso a la familia. Con símbolos se escribiría : $F(5,3) = 0.54$

A continuación se presenta un resumen de la notación y las propiedades de las frecuencias.

NOTACION Y PROPIEDADES

n = número total de elementos de la muestra

n_{ij} = número de elementos de la muestra que pertenecen en forma simultánea a las categorías X_i y Y_j

$n_{i.}$ = número de elementos de la muestra que pertenecen a la categoría X_i.

$n_{.j}$ = número de elementos de la muestra que pertenecen a la categoría Y_j.

f_{ij} = fracción (o porcentaje) del total de elementos de la muestra que pertenecen simultáneamente a las categorías X_i y Y_j

$$= \frac{n_{ij}}{n}$$

$f_{i.}$ = fracción (o porcentaje) del total de elementos de la muestra, que pertenecen a la categoría X_i.

$$= \frac{n_{i.}}{n}$$

$f_{.j}$ = fracción (o porcentaje) del total de elementos de la muestra que pertenecen a la categoría Y_j.

$$= \frac{n_{.j}}{n}$$

$N(x,y)$ = número de elementos cuya característica X es menor o igual que x, y su característica Y es menor o igual que y.

$F(x,y)$ = fracción (o porcentaje) de elementos para los cuales $X \leq x$ y $Y \leq y$.

$$= \frac{N(x,y)}{n}$$

Como puede deducirse del ejemplo 3.1, se cumplen las siguientes propiedades:

1. $\sum_{i=1}^{m}\sum_{j=1}^{s} n_{ij} = n_{11} + n_{12} + ... + n_{1s} + n_{21} + n_{22} + ... + n_{2s} + ... + n_{m1} + n_{m2} + n_{ms} = n$

2. $\sum_{i=1}^{m} n_{ij} = n_{1j} + n_{2j} + ... + n_{mj} = n_{.j}$

3. $\sum_{j=1}^{s} n_{ij} = n_{i1} + n_{i2} + ... + n_{is} = n_{i.}$

4. $\sum_{i=1}^{m} n_{i.} = n$

5. $\sum_{j=1}^{s} n_{.j} = n$

De las anteriores propiedades, al dividir por *"n"* se obtiene para las frecuencias relativas:

6. $\sum_{i=1}^{m}\sum_{j=1}^{s} f_{ij} = 1.00$

7. $\sum_{i=1}^{m} f_{ij} = f_{.j}$

8. $\sum_{j=1}^{s} f_{ij} = f_{i.}$

9. $\sum_{i=1}^{m} f_{i.} = 1.00$

10. $\sum_{j=1}^{s} f_{.j} = 1.00$

Para las frecuencias acumuladas puede escribirse:

11. Si $X_1 < X_2 < ... < X_m$
$Y_1 < Y_2 < ... < Y_s$

Entonces:

Si $x < X_1$, $y < Y_1 \Rightarrow F(x,y) = 0$

Si $x \geq X_m$, $y \geq Y_s \Rightarrow F(x,y) = 1.00$

12. Si $x < x^*$; $y < y^* \Rightarrow F(x,y) \leq F(x^*, y^*)$

OBSERVACIÓN

A la distribución de frecuencias $\{(X_i, Y_j, f_{ij})\}$ se le conoce como "función empírica de distribución conjunta de frecuencias de las variables *X* e *Y*".

A la distribución de frecuencias $\{(x,y, F(x,y)\}$ se le conoce como "función empírica de distribución acumulada de las variables *X* e *Y*".

3.2.2 Caso en el cual ambas variables (*X,Y*) son continuas

En este caso, las categorías a considerar para cada variable están representadas por intervalos de clase, que se construyen de la forma propuesta en el capítulo 2.

Casi todos los conceptos desarrollados para la situación en que ambas variables son discretas son válidos aquí, incluyendo las propiedades de las distribuciones de frecuencia. Sin embargo, es particular en el tratamiento de variables continuas, por su naturaleza, el concepto de función empírica de densidad conjunta de las variables *X* e *Y*. Esta temática se desarrolla a través del ejemplo que se presenta a continuación:

Ejemplo 3.2

En un estudio realizado en la región del Omait en el cual la población de interés estaba constituida por las fincas que cultivan maíz, se tomó al azar una muestra de 200 fincas de las cuales se registra las variables: área cultivada, *X*, en hectáreas y producción anual de maíz, *Y*, en toneladas.

Con base en los 200 datos, se construyó los siguientes intervalos de clase:

X: Área cultivada (Ha)

X_1 : (0;10]; X_2 : (10;40]; X_3 : (40;90]; X_4 : (90;150]

Y: Producción anual de maíz (ton)

Y_1 : (0;25] ; Y_2 : (25;60] ; Y_3 : (60;180] ; Y_4 : (180;250] ; Y_5 : (250;350]

De acuerdo con los anteriores intervalos de clase se construyó el siguiente cuadro de frecuencias:

CUADRO 3.6

DISTRIBUCION CONJUNTA DE FRECUENCIAS ABSOLUTAS PARA EL AREA CULTIVADA (*X*) Y LA PRODUCCION ANUAL DE MAIZ (*Y*)

Y / X	(0 ; 25] Y_1	(25 ; 60] Y_2	(60 ; 180] Y_3	(180 ; 250] Y_4	(250 ; 350] Y_5	
(0 ; 10] X_1	34	30	14	2	0	80
(10 ; 40] X_2	23	12	20*	4	1	60*
(40 ; 90] X_3	13	8	24	4	1	50
(90 ; 150] X_4	0	0	2	5	3	10
	70	50	60	15	5*	200

La interpretación de los valores de este cuadro, es completamente análoga a la presentada para variables discretas, así pues:

Hay en la muestra 20 fincas cuya área cultivada está entre 10 y 40 hectáreas y cuya producción anual de maíz está entre 60 Ton. y 180 Ton. Usando la notación se escribiría $n_{23} = 20$.

Hay en la muestra 60 fincas con un área cultivada de maíz en el intervalo 10 hectáreas a 40 hectáreas, es decir $n_{2.} = 60$.

Hay 5 fincas que producen al año entre 250 y 350 Ton. de maíz, es decir $n_{.5} = 5$

CUADRO 3.7

DISTRIBUCION CONJUNTA DE FRECUENCIAS RELATIVAS PARA EL AREA CULTIVADA (*X*) Y LA PRODUCCION ANUAL DE MAIZ (*Y*)

Y / X	(0 ; 25] Y_1	(25 ; 60] Y_2	(60 ; 180] Y_3	(180 ; 250] Y_4	(250 ; 350] Y_5	
(0 ; 10] X_1	0.170	0.150	0.070	0.010	0	0.40
(10 ; 40] X_2	0.115	0.060	0.100*	0.020	0.005	0.30*
(40 ; 90] X_3	0.065	0.040	0.120	0.020	0.005	0.25
(90 ; 150] X_4	0	0	0.010	0.025	0.015	0.05
	0.35	0.25	0.30	0.075	0.025*	1.00

Los valores f_{ij} de este cuadro se obtienen expresando el número de datos, como una fracción (o porcentaje) del total de 200 datos, es decir:

$$f_{ij} = \frac{n_{ij}}{n}$$

La interpretación de las cifras del cuadro 3.7, es la de un porcentaje, de esta manera:

- 0.100 indica que el 10% de las fincas tienen área cultivada de maíz entre 10 y 40 hectáreas y a la vez tienen producción anual entre 60 y 180 Ton. $f_{23} = 0.100$.

- 0.30 indica que el 30% de las fincas de la muestra tienen área cultivada de maíz entre 10 y 40 hectáreas, es decir $f_{2.} = 0.30$

- 0.025 indica que el 2.5% de las fincas producen al año entre 250 Ton. y 350 Ton. de maíz, o sea $f_{.5} = 0.025$.

Observe que de igual manera que en el caso discreto, pueden construirse las distribuciones marginales tanto para el área cultivada (*X*), como para la producción anual de maíz (*Y*).

Los porcentajes o fracciones que aparecen en el cuadro 3.7, no son directamente comparables puesto que los intervalos de clase construidos tanto para *X* como para *Y* son de longitudes o anchos distintos, en realidad podría decirse que las regiones que están determinadas por la doble partición:

$(X_1 \cap Y_1),(X_1 \cap Y_2)$, ... , $(X_1 \cap Y_5)$, $(X_2 \cap Y_1)$, ... , $(X_2 \cap Y_5)$, ... , $(X_4 \cap Y_5)$

tienen **áreas** diferentes. De manera análoga como se resolvió la situación en el caso unidimensional, definiendo el concepto de densidad por unidad de intervalo, se plantea la estandarización de las frecuencias relativas definiendo el concepto de **densidad por unidad de área,** de esta forma si se denota por:

A_{ij} = área de la región determinada por $(X_i \cap Y_j)$

se puede definir la densidad:

$$f^*_{ij} = \frac{f_{ij}}{A_{ij}} \quad \text{para la región } X_i \cap Y_j$$

con el supuesto de que los datos en cada región están uniformemente distribuídos.

Al definir f^*_{ij} para cualquier punto del plano *X* - *Y*, se obtiene la llamada **función empírica de densidad conjunta de *X* e *Y*.**
Para el ejemplo 3.2, las áreas de las distintas regiones definidas por los intervalos de clase en *X* e *Y* se muestran en el cuadro 3.8.

Los valores del cuadro se calcularon con base en los productos de las longitudes de los intervalos correspondientes. Dado que *X* está en hectáreas e *Y* está dado en Ton., las unidades del área calculada son hectáreas x toneladas.

CUADRO 3.8

AREAS DE LAS REGIONES DEFINIDAS SOBRE EL PLANO *X-Y*, POR LOS INTERVALOS DE CLASE RESPECTIVOS.
(A_{ij}) *(Hectáreas x Toneladas)*

Y / X	(0 ; 25] Y_1	(25 ; 60] Y_2	(60 ; 180] Y_3	(180 ; 250] Y_4	(250 ; 350] Y_5
(0 ; 10] X_1	250	350	1200	700	1000
(10 ; 40] X_2	750	1050	3600	2100	3000
(40 ; 90] X_3	1250	1750	6000	3500	5000
(90 ; 150] X_4	1500	2100	7200	4200	6000

Con base en los cuadros 3.7 y 3.8, puede calcularse la densidad:

$$f^*_{ij} = \frac{f_{ij}}{A_{ij}}$$

lo cual da origen al cuadro 3.9, en el cual se presenta la función de densidad conjunta, en la cual se expresa *(Hectáreas)*$^{-1}$ x *(Toneladas)*$^{-1}$.

CUADRO 3.9

FUNCION EMPIRICA DE DENSIDAD CONJUNTA PARA LAS VARIABLES AREA CULTIVADA (Ha) Y PRODUCCION ANUAL DE MAIZ (Ton), EN LAS FINCAS DE LA REGION DE OMAIT.

$f^*(x, y)$ en $F^{-1}{}_a$ x Ton^{-1}

Y X	(0 ; 25] Y_1	(25 ; 60] Y_2	(60 ; 180] Y_3	(180 ; 250] Y_4	(250 ; 350] Y_5
(0 ; 10] X_1	0.00068000	0.00042857	0.00005833	0.00001428	0
(10 ; 40] X_2	0.00015333	0.00005714	0.00002777	0.00000952	0.00000166
(40 ; 90] X_3	0.00005200	0.00002285	0.00002000	0.00000571	0.00000100
(90 ; 150] X_4	0	0	0.00000138	0.00000595	0.00000250

Lógicamente en cualquier región distinta a la cubierta por el cuadro 3.9, $f^*(x,y) = 0$.

La representación gráfica de la función empírica de densidad conjunta, aparece en la figura 3.3, la cual es una ampliación del concepto de histograma, con la diferencia que en lugar de hablarse de área se habla de volumen.

Si se calcula el volumen del paralelepípedo que está sobre la región definida por $(X_i \cap Y_j)$, se obtiene:

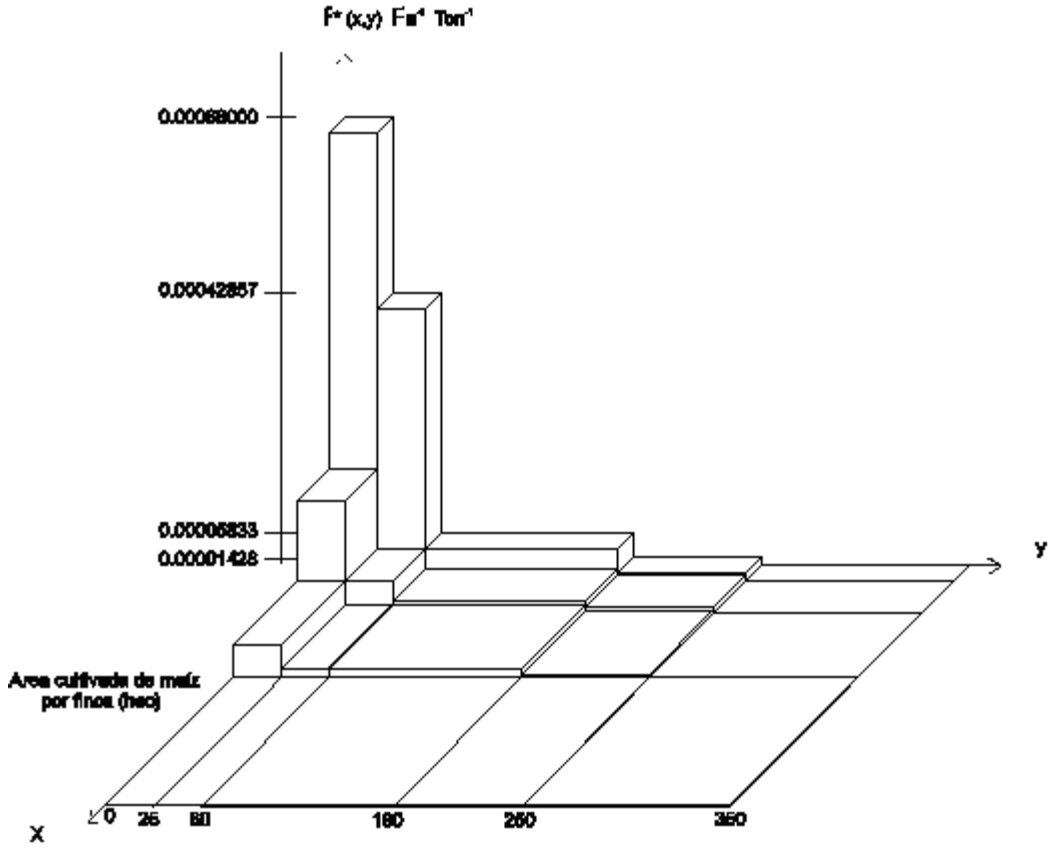


FIG. 3.3. Función empírica de densidad conjunta para las variables "área cultivada" y "producción anual de maíz"

$$V = \text{área de la base } x \text{ altura}$$
$$V_{ij} = A_{ij} \quad x \quad f^*_{ij}$$

Dado que: $f^*_{ij} = \dfrac{f_{ij}}{A_{ij}}$, entonces

$$V_{ij} = f_{ij}$$

Lo cual significa que el volumen de un prisma representa la frecuencia relativa (porcentaje de datos) que pertenecen a la región definida por la base del mismo, por tal razón al calcular el volumen total del gráfico debe arrojar como resultado 100%

Aplicando estos conceptos, puede estimarse el porcentaje de datos que pertenecen a cualquier región del plano *X - Y*, tan sólo calculando el volumen que se levanta sobre la mencionada región como se presenta en el siguiente ejemplo.

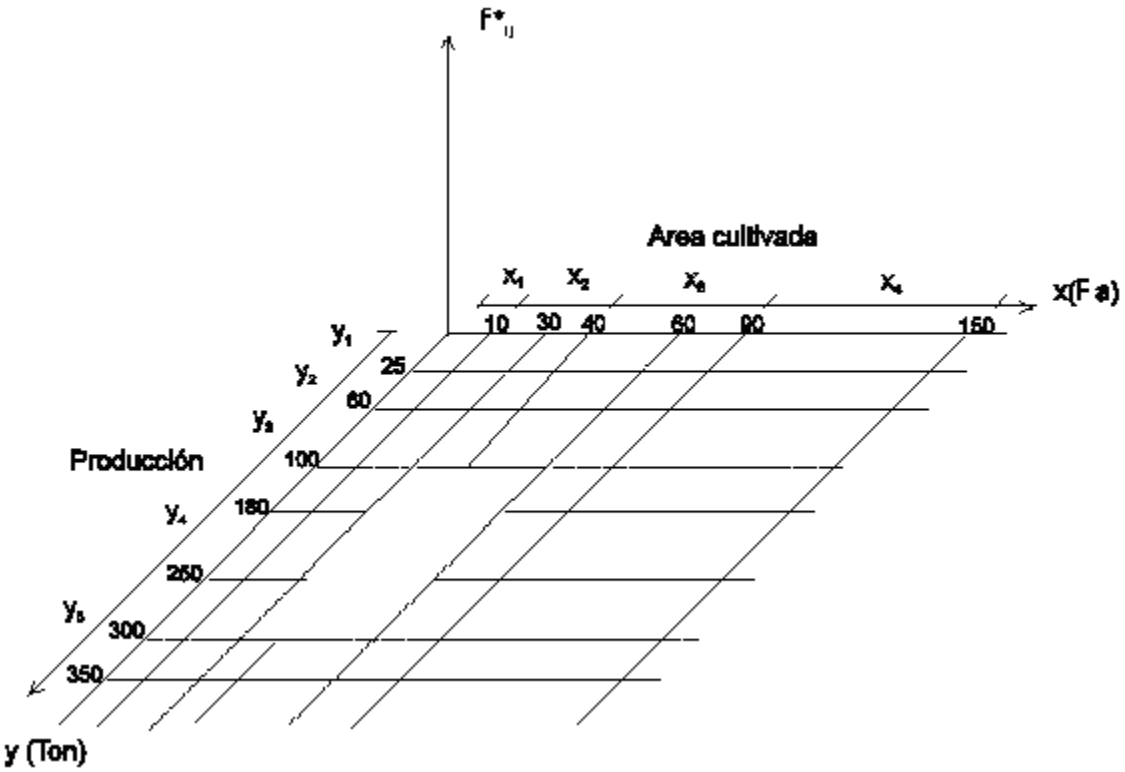
Ejemplo 3.3

Con base en la información proporcionada por el ejemplo 3.2, estime el porcentaje de fincas que tienen áreas de cultivo de maíz entre 30 Ha y 60 Ha y producen anualmente entre 100 Ton. y 300 Ton.

La solución al problema planteado consiste en calcular el volumen del gráfico de la figura 3.3, sobre la región pedida que aparece sombreada en el siguiente esquema, donde se muestra que la región sombreada es la unión de seis "pedazos" que pertenecen a regiones distintas de las establecidas en el ejemplo anterior y, por lo tanto, pueden tener alturas (f^*_{ij}) diferentes, en consecuencia debe hallarse cada uno de los volúmenes pertinentes y luego realizar la suma, por tal razón en el esquema siguiente aparecen delimitadas las distintas regiones que se deben considerar; de esta manera:

R_{ij} = área del "pedazo" de la región sombreada que pertenece a la región definida por $(X_i \cap Y_j)$, la cual tiene densidad f^*_{ij}

Por tanto el volumen total sobre la zona sombreada y que corresponde a la solución del problema es:



f(región sombreada) = V(R_{ij} es la región sombreada que esta incluida en $(X_i \cap Y_j)$ y que por lo tanto tiene densidad constante f^*_{ij}) = $R_{23} \cdot f^*_{23} + R_{33} \cdot f^*_{33} + R_{24} \cdot f^*_{24} +$

$$R_{34} \cdot f^*_{34} + R_{25} \cdot f^*_{25} + R_{35} f^*_{35}$$

El área de las regiones requeridas se presenta a continuación:

Región	R_{23}	R_{33}	R_{24}	R_{34}	R_{25}	R_{35}
Area = Ha x Ton	800	1600	700	1400	500	1000

Por tanto el porcentaje de fincas con áreas cultivada de maíz entre 30 Fa. y 60 Fa. y con producción anual entre 100 Ton. y 300 Ton., está dado por

= 800 x 0.0000277 + 1600 x 0.0000200 + 700 x 0.0000095 + 1400 x 0.0000057 + 500 x 0.0000016 + 1000 x 0.00000100 $\approx$ 0.08 $\equiv$ 8%

Puede definirse la **función empírica de distribución conjunta acumulada** para las variables *X* e *Y*, que se representa por *F(x,y)* y se obtiene del cálculo del volumen correspondiente a la región comprendida por $X \leq x$ e $Y \leq y$, haciendo las consideraciones de que el punto *(x,y)* pertenezca a cada una de las distintas regiones que determinan los $(X_i \cap Y_j)$, de forma que si $(x,y) \in (X_2 \cap Y_3)$, entonces

F(x,y) = fracción del lote de datos que satisfacen que $X \leq x$ e $Y \leq y$

$$F(x,y) = R_{11} \cdot f^*_{11} + R_{12} \cdot f^*_{12} + R_{13} \cdot f^*_{13} + R_{21} \cdot f^*_{21} + R_{22} \cdot f^*_{22} + R_{23} f^*_{23}$$
$$= 250 \text{ x } 0.00068 + 350 \text{ x } 0.0004286 + (y\text{-}60) \text{ x } 0.0000583 \text{ x } 10 + (x\text{-}10) \text{ x } 25 \text{ x } 0.0001533 + (x\text{-}10) \text{ x } 35 \text{ x } 0.0000571 + (x\text{-}10) \text{ x } (y\text{-}60) \text{ x } 0.0000277$$

Este procedimiento se repetiría para cada una de las regiones $(X_i \cap Y_j)$

3.2.3 Caso en el cual una variable es discreta y la otra es continua.

Supóngase que *X* es una variable discreta y *Y* es continua; en este caso al organizar la muestra bruta, las categorías para *X* las constituyen los valores distintos que toma la variable, en cambio para *Y* se deben construir intervalos de clase. De esta manera se pueden clasificar y contar los datos de la muestra para dar origen a un cuadro que representa la distribución conjunta de frecuencias relativas para (X_i, Y_j); también puede expresarse las frecuencias absolutas como una fracción (o porcentaje) del total de elementos para dar origen a un cuadro de frecuencias relativas para (X_i,Y_j). Dado que la variable *Y* es continua, tiene sentido hablar de la función empírica de densidad de *Y*, más no de *X*; por tal razón, estrictamente hablando no sería muy adecuado referirse a la función empírica de densidad conjunta de (*X*,*Y*), puesto que *X* es discreta; no obstante lo anterior y con el propósito de no usar nuevos términos para hacer referencia a conceptos similares, se va a usar el nombre de función empírica de densidad conjunta $f^*(x,y)$, pero haciendo la precisión de su significado y su forma de operación, para ello se presenta un ejemplo a continuación.

Ejemplo 3.4

Se tomó una muestra de 500 hogares en los cuales se observó las características:

X : número de personas que constituyen el hogar
Y : ingreso del hogar (en miles de pesos)

Los valores distintos encontrados para la variable X fueron:

$X_1 = 1; \quad X_2 = 2; \quad X_3 = 3; \quad X_4 = 5$

Para la variable Y, ingresos del hogar (en miles de pesos) se construyeron los siguientes intervalos de clase:

Y_1 : (50;75] ; Y_2 : (75;125] ; Y_3 : (125;200] ; Y_4 : (200;300] ; Y_5 : (300;550]

Con base en la categorización anterior se clasificaron los datos y al realizar el conteo se construyó el siguiente cuadro.

CUADRO 3.10

DISTRIBUCION CONJUNTA DE FRECUENCIAS ABSOLUTAS DEL NUMERO DE PERSONAS (X) Y EL INGRESO DEL HOGAR (Y).

Y / X	(50 ; 75] Y_1	(75 ;125] Y_2	(125 ; 200] Y_3	(200 ; 300] Y_4	(300 ; 550] Y_5	
X_1=1	36	15	12	9	3	75
X_2=2	38	20	23	14	5	100
X_3=3	86	60	25	22	7	200
X_4=5	15	30	40	30	10	125
	175	125	100	75	25	500

Al expresar las frecuencias absolutas como una fracción con respecto al número total de elementos obtenemos el cuadro 3.11.

Dado que en la pareja (X,Y), sólo Y es una variable continua, la convenida función empírica de densidad conjunta, resulta de estandarizar la frecuencia relativa f_{ij} por unidad de intervalo de Y_j

$$f_{ij}^{*} = \frac{f_{ij}}{C_j}$$

donde C_j = longitud del intervalo Y_j

CUADRO 3.11

DISTRIBUCION CONJUNTA DE FRECUENCIAS RELATIVAS DEL NUMERO DE PERSONAS (*X*) Y EL INGRESO DEL HOGAR (*Y*).

Y / X	(50 ; 75] Y_1	(75 ; 125] Y_2	(125 ; 200] Y_3	(200 ; 300] Y_4	(300 ; 550] Y_5	
X_1=1	0.072	0.030	0.024	0.018	0.006	0.15
X_2=2	0.076	0.040	0.046	0.028	0.010	0.20
X_3=3	0.172	0.120	0.050	0.044	0.014	0.40
X_4=5	0.030	0.060	0.080	0.060	0.020	0.25
	0.35	0.25	0.20	0.15	0.05	1.00

De esta manera f^*_{ij} es una densidad por unidad lineal y no por área.

Con este proceso se da origen al cuadro 3.12, donde la función empírica de densidad conjunta de *X* e *Y* puede definirse como:

$$f^*(x,y) = \begin{cases} f^*_{ij} & \text{si } (x,y) \in (X_i \cap Y_j) \quad , \; i = 1, 2, ..., m \\ & \qquad\qquad\qquad\qquad\quad j = 1, 2, ..., s \\ 0 & \text{en cualquier otra parte} \end{cases}$$

CUADRO 3.12

FUNCION EMPIRICA DE DENSIDAD CONJUNTA DE LAS VARIABLES NUMERO DE PERSONAS (*X*) Y EL INGRESO DEL HOGAR (*Y*).

$f^*(x,y)$ *en (miles de pesos)*$^{-1}$

Y / X	(50 ; 75] Y_1	(75 ; 125] Y_2	(125 ; 200] Y_3	(200 ; 300] Y_4	(300 ; 550] Y_5
X_1=1	0.00288000	0.00060000	0.00032000	0.00018000	0.00002400
X_2=2	0.00304000	0.00080000	0.00061333	0.00028000	0.00004000
X_3=3	0.00688000	0.00240000	0.00066666	0.00044000	0.00005600
X_4=5	0.00120000	0.00120000	0.00106666	0.00060000	0.00008000

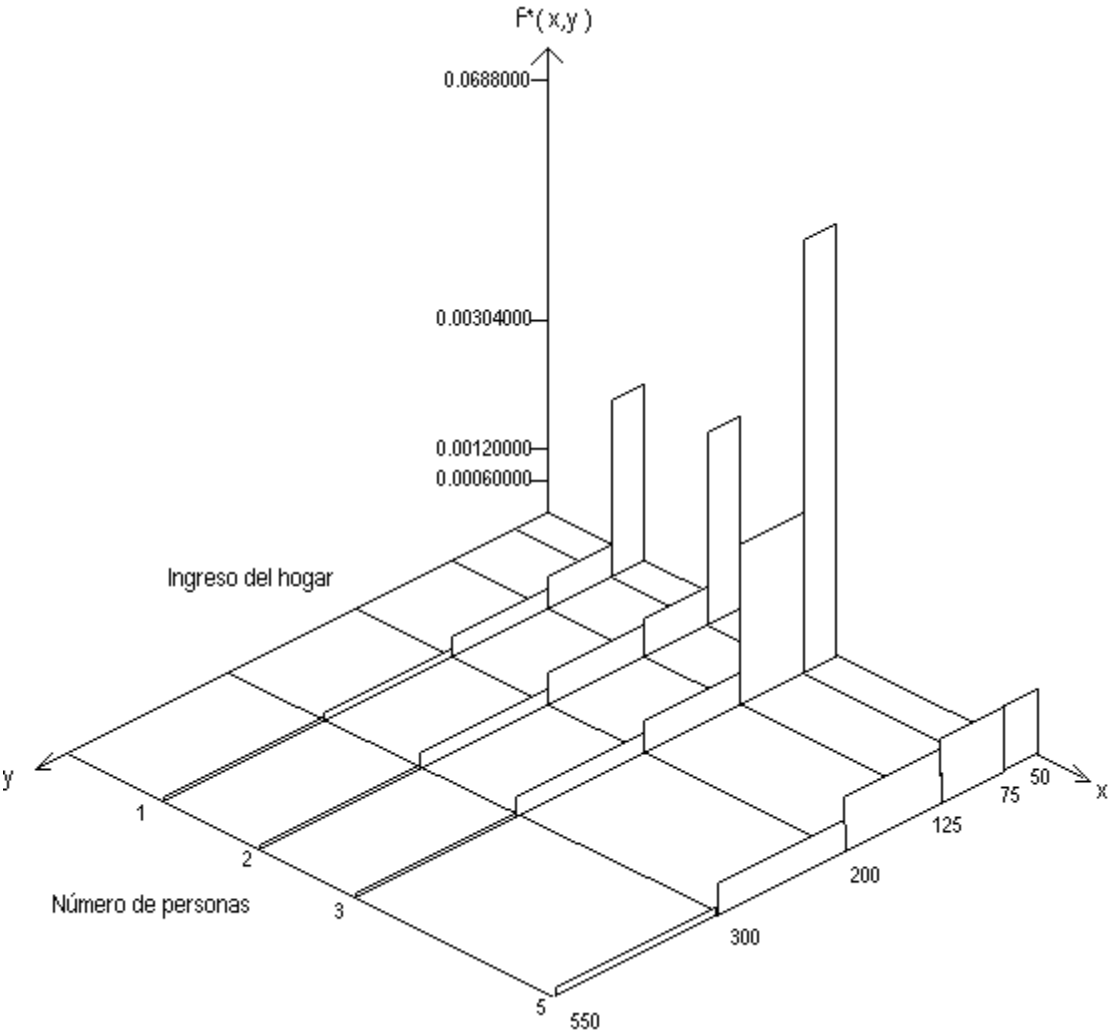


Fig. 3.4 : Función empírica de densidad conjunta de (X,Y) cuando X es discreta y Y continua.

Como es lógico la suma de las áreas de todas las "láminas" es 1.0 (ó 100%)

Ejemplo 3.5

Con base en la función empírica de densidad conjunta para el número de personas por hogar y el ingreso del hogar estime:

a) El porcentaje de hogares que tienen 3 personas e ingresos entre \$90.000 y \$275.000.
Observando la figura 3.4, se trata de calcular el área comprendida entre $Y = 90$ y $Y = 275$ en la "lámina" correspondiente a $X = 3$.

El porcentaje pedido es:

$$f^*_{32}\,(125\text{-}90) + f^*_{33}\,(200\text{-}125) + f^*_{34}\,(275\text{-}200)$$
$$= 0.00240(35) + 0.000666(75) + 0.000440(75)$$
$$= 0.167$$

lo cual significa que aproximadamente el 16.7% de los hogares están compuestos por 3 personas y tienen ingresos entre \$90.000 y \$275.000.

b) El porcentaje de hogares con 2 ó 3 personas y con ingresos entre \$90.000 y \$275.000.
Ahora deben calcularse las áreas comprendidas entre $Y = 90$ y $Y = 275$ en las láminas correspondientes a $X = 2$ *y* $X = 3$ y deben sumarse:

para $X = 2$ el área es:

$$f^*_{22}\,(125\text{-}90) + f^*_{23}\,(200\text{-}125) + f^*_{24}\,(275\text{-}200)$$
$$= 0.000800(35) + 0.000613(75) + 0.00028(75)$$
$$= 0.095$$

para $X = 3$ el área es la hallada en a), es decir $= 0.167$

El porcentaje pedido es 9.5% + 16.7% = 26.2%

3.3 DISTRIBUCIONES CONDICIONALES DE FRECUENCIA

En muchas ocasiones es de interés la distribución de frecuencias relativas de una característica, pero sólo para los elementos de la muestra que satisfacen cierta condición, por ejemplo, la distribución de la característica "estatura" para las personas de la muestra que tienen "peso corporal" entre 60 kg, y 70 kg; la distribución del "ingreso familiar" para los hogares que están constituidos por 4 personas; la distribución del área cultivada de maíz" para las fincas con "producción anual entre 70 Ton. y 100 Ton.; o la distribución de frecuencias de la "producción anual de maíz" para las fincas con "área cultivada" entre 30 y 40 hectáreas; la distribución de frecuencias de la opinión sobre la legalización del consumo de marihuana para los votantes potenciales con edades entre 20 y 30 años; la distribución de frecuencia de padecer o no cierta enfermedad para los fumadores con hábito desde más de 10 años.
Cuando se hace referencia, como en las situaciones anteriores, a la distribución de una variable para los elementos de una muestra que satisfacen cierta condición se le llamará **distribución condicional de frecuencias.** La condición puede ser de cualquier naturaleza: en general, si *"C"* es el conjunto de elementos de la muestra, que satisfacen la condición *"C"*, entonces:

f(A/C) representa el porcentaje (o fracción) de los elementos de *C* que pertenecen al conjunto *A* y se lee "frecuencia relativa de *A* dado *C*".

Ejemplo 3.6

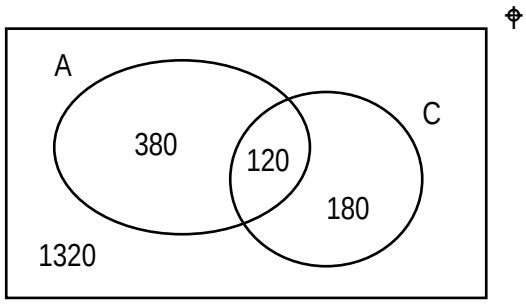
De una muestra de 2.000 viviendas se observó la tenencia de servicios de agua y energía :

Ω = es el conjunto de viviendas observadas en la muestra
A = es el conjunto de viviendas con servicio de agua potable.
C = es el conjunto de viviendas con servicio de energía eléctrica.

y el número de elementos de cada conjunto es:

$n(\Omega)$ = 2.000 viviendas en la muestra
$n(A)$ = 500 viviendas con agua
$n(C)$ = 300 viviendas con energía eléctrica
$n(A \cap C)$ = 120 viviendas con agua y energía eléctrica

El esquema que se presenta a continuación muestra los distintos conjuntos que resultan en la muestra Ω y el número de sus respectivos elementos.



con la notación que se ha presentado, puede deducirse del esquema, lo siguiente:

$f(A/C) = \frac{120}{300} = 0.40$, lo cual significa que de las viviendas con energía, el 40% tienen agua potable.

$f(C/A) = \frac{120}{500} = 0.24$, lo cual significa, que de las viviendas con agua potable, el 24% tienen energía eléctrica.

$f(A \cap C) = \frac{120}{2000} = 0.06$, lo cual significa que de todas las viviendas observadas, el 6% tienen simultáneamente agua y energía. (Note que esta frecuencia no es condicional).

Si se usa la notación de complemento de conjuntos:

$\overline{A}$: es el conjunto de viviendas que no tienen servicio de agua potable.

$\overline{C}$: es el conjunto de elementos que no tienen energía eléctrica.

Puede calcularse:

$f(A/\overline{C}) = \frac{380}{1700} = 0.2235$; significa que de las viviendas que no tienen energía eléctrica, el 22.35% de ellas, tienen agua potable.

$f(\overline{A}/C) = \frac{180}{300} = 0.60$; significa que de las viviendas que tienen energía eléctrica, el 60% no tienen servicio de agua potable.

$f(C/\overline{A}) = \frac{180}{1500} = 0.12$; significa que de las viviendas que no tienen agua, el 12% de ellas tienen energía.

$f(\overline{C}/\overline{A}) = \frac{1320}{1500} = 0.88$; significa que de las viviendas que no tienen agua, el 88% de ellas no tienen energía.

$f(\overline{A} \cap \overline{C}) = \frac{1320}{2000} = 0.66$; significa que de todas las viviendas observadas, el 66% no tienen agua ni energía. (No es una frecuencia condicional)

$f(A) = \frac{500}{2000} = 0.25$; significa que de todas las viviendas observadas el 25% tienen servicio de agua. (No es una frecuencia condicional).

$f(C) = \frac{300}{2000} = 0.15$; de todas las viviendas observadas, el 15% tienen servicio de energía eléctrica.

Del ejemplo anterior puede obtenerse una definición para la frecuencia condicional

de *A* dado *C*, así:

$$f(A/C) = \frac{n(A \cap C)}{n(C)}$$

si se divide el numerador y el denominador por el número total de elementos de la muestra $n(\Omega)$, se obtiene:

$$f(A/C) = \frac{f(A \cap C)}{f(C)}$$

Expresión que permite obtener la frecuencia condicional, como el cociente de frecuencias no condicionales.

En forma análoga, cuando se tiene la distribución conjunta de (X_i, Y_j), puede escribirse:

$$f(X_i / Y_j) = \frac{n(X_i \cap Y_j)}{n(Y_j)} = \frac{n_{ij}}{n_{.j}} = \frac{n_{ij}/n}{n_{.j}/n}$$
$$= \frac{f_{ij}}{f_{.j}} = \frac{f(X_i \cap Y_j)}{f(Y_j)}$$

También:

$$f(Y_j / X_i) = \frac{f_{ij}}{f_{i.}} = \frac{f(X_i \cap Y_j)}{f(X_i)}$$

Como puede observarse de la definición de $f(X_i/Y_j)$ se satisface que:

i) $f(X_i/Y_j) \geq 0$ para todo *i* y *j*

ii) $\sum_{i=1}^{m} f(X_i / Y_j) = 1$

$\{X_i , f(X_i/Y_j)\}$ constituye la distribución condicional de X , $i = 1,2,...,m$, dado Y_j.

$\{Y_j , f(Y_j/X_i)\}$ es la distribución condicional de Y, dado X_i , $j = 1,2,...,s$

Las distribuciones condicionales de frecuencias, satisfacen todas las propiedades definidas para las distribuciones de frecuencias relativas, por tanto en el caso de

variables continuas, tiene sentido referirse a la **función empírica de densidad condicional de *Y* dado *X* , *f*(Y/X),*** que se definirá como:

$$f^*\left(Y_j / X_i\right) = \frac{f\left(Y_j / X_i\right)}{C_{Y_j}}$$

Donde C_{Y_j} = la longitud del intervalo *j* de *Y*.

Ejemplo 3.7

Haciendo referencia al ejemplo 3.2, en el cual se observa una muestra de 200 fincas, las variables área cultivada de maíz (*X*) en Ha, y producción anual (*Y*) en Ton. se presenta a continuación la distribución conjunta de frecuencias absolutas

Y X	(0 ; 25] Y_1	(25 ; 60] Y_2	(60 ; 180] Y_3	(180 ; 250] Y_4	(250 ; 350] Y_5	
(0 ; 10] X_1	34	30	14	2	0	80
(10 ; 40] X_2	23	12	20	4	1	60
(40 ; 90] X_3	13	8	24	4	1	50
(90 ; 150] X_4	0	0	2	5	3	10
	70	50	60	15	5	200

a) Construir la distribución condicional del área cultivada, para las fincas con producción anual entre (60; 180), dicha distribución se denota por $\{X_i ; f(X_i/Y_3)\}$

Area Cultivada (X_i)	$f(X_i/Y_3)$
X_1: (0 ; 10]	14/60
X_2 : (10 ; 40]	20/60
X_3 : (40 ; 90]	24/60
X_4 : (90 ; 150]	2/60
	1.00

En el cuadro anterior:

$f(X_2/Y_3)$ = 20/60 = 0.33 significa que de las fincas que producen entre 60 y 180 Ton. de maíz al año, el 33% de ellas tienen área cultivada entre 10 Ha. y 40 Ha.

b) Construir la función empírica de densidad condicional del área cultivada, para las fincas con producción anual entre 60 Ton. y 180 Ton.

$$f^*(X_i/Y_3) = \frac{f(X_i/Y_3)}{C_{X_i}}$$

como ejemplo:

$$f^*(X_1/Y_3) = \frac{f(X_1/Y_3)}{C_{X_1}} = \frac{14/60}{10} = \frac{14}{600} = 0.023$$

De esta manera puede definirse:

$$f^*(x/y_3) = \begin{cases} 0.0000 & \text{si } x < 0 \text{ ó } x > 150 \\ 0.0233 & \text{si } x \in (0;10] \\ 0.0110 & \text{si } x \in (10;40] \\ 0.0080 & \text{si } x \in (40;90] \\ 0.0006 & \text{si } x \in (90;150] \end{cases}$$

c) Calcule qué porcentaje de las fincas que producen anualmente entre 60 y 180 Ton. de maíz tienen áreas cultivadas entre 18 Ha. y 70 Ha.

$$f(18 \le X \le 70/Y_3) = f^*(X_2/Y_3)(40-18) + f^*(X_3/Y_3)(70-40)$$
$$= 0.011 \text{ x } 22 + 0.008 \text{ x } 30$$
$$= 0.482 = 48.2\%$$

d) Calcule e interprete $f(X_1/Y_3)$, $f(Y_3/X_1)$, $f(X_1 \cap Y_3)$

$f(X_1/Y_3) = \frac{14}{60} = 0.233$, significa que de las fincas que producen anualmente entre 60 y 180 Ton. de maíz, el 23,3% de ellas, tienen área cultivada entre 0 y 10 hectáreas.

$f(Y_3/X_1) = \frac{14}{80} = 0.175$, significa que de las fincas con área cultivada de maíz entre 0 y 10 hectáreas, el 17.5% de ellas producen anualmente entre 60 y 180 Ton. de maíz.

$f(X_1 \cap Y_3) = \frac{14}{200} = 0.07$, significa que de todas las fincas de la muestra, el 7% tienen área cultivada entre 0 y 10 Ha, y producen al año entre 60 y 180 Ton. de maíz.

De la definición de la distribución condicional

$$f\left(X_i / Y_j\right) = \frac{f\left(X_i \cap Y_j\right)}{f\left(Y_j\right)}$$

ó

$$f\left(Y_j / X_i\right) = \frac{f\left(X_i \cap Y_j\right)}{f\left(X_i\right)}$$

puede deducirse la llamada **regla de multiplicación,** como:

$$f(X_i \cap Y_j) = f(Y_j)\, f(X_i/Y_j)$$

ó

$$f(X_i \cap Y_j) = f(X_i)\, .\, (Y_j/X_i)$$

En resumen la regla de la multiplicación expresa que la distribución conjunta de frecuencias relativas puede escribirse como el producto de la distribución marginal de una de las variables por la condicional de la otra.

3.3.1 Algunos casos de interpretación equivocada de frecuencias condicionales.

En ocasiones los medios de comunicación corrientes y aún la literatura científica cometen errores de interpretación, sin mala intención en la mayoría de los casos.

A continuación se presentan varias de estas situaciones, con el propósito de que se reflexione un poco al respecto y se lea con mucha prevención la literatura que hace referencia a este tipo de cifras.

1. En la población de "Polulandia" el 50% de las consultas son por enfermedades respiratorias, en cierto período, lo cual permite inducir que existen precarias condiciones ambientales que afectan a las personas en su aparato respiratorio.

- Nótese que en el planteamiento anterior no hay ninguna afirmación que proporcione información acerca de si las consultas son "muchas" o son "pocas" con respecto al número de habitantes de la población. La afirmación sobre el 50%, es válida aun en el caso en que en una población de un millón de habitantes se hayan hecho dos (2) consultas al año, una de las cuales por afecciones respiratorias. En este caso se estaría confundiendo *f(R/C)* con *f(R)* donde: *f(R/C)* = porcentaje que representan las consultas por enfermedades respiratorias con respecto al número total de consultas realizadas y *f(R)* = porcentaje de consultas por enfermedades respiratorias, con respecto a toda la población.

2. Una encuesta realizada por un periódico entre los intoxicados que habían asistido a una boda, mostró que el 90% de ellos había comido pollo. Esto es una clara indicación de la fuente de contagio.

   De nuevo en este caso, no se presenta información sobre si los intoxicados son "muchos" o "pocos" comparados con todos los que comieron pollo.
   La frecuencia que menciona el enunciado es:

   $$f(P/I) = 0.90$$

   Sería de más valor comparar el porcentaje de intoxicados entre los que comieron pollo con el porcentaje de intoxicados que no comieron pollo, es decir:

   $$f(I/P) \text{ con } f(I/\overline{P})$$

   Aunque tampoco sería del todo concluyente, véase por qué : supóngase que en el peor de los casos:

   $$f(I/P) = 100\% \quad \text{y} \quad f(I/\overline{P}) = 0\%$$

   Es decir, todos los que comieron pollo se intoxicaron y de los que no comieron pollo ninguno se intoxicó, aun así, no puede atribuirse la culpa al pollo con absoluta seguridad, puesto que pudo pasar lo siguiente:
   Todos los que comieron pollo, tomaron sobremesa y los que no comieron pollo no tomaron sobremesa y, puede haber sido ésta la causa, puesto que en estas circunstancias también se obtienen los mismos resultados numéricos.
   Esta situación permite visualizar que las asociaciones estadísticas entre eventos no guardan necesariamente una relación de causa a efecto.

3. Una encuesta entre prostitutas realizada en Cali mostró que un elevado porcentaje de ellas, más del 80%, habían nacido en el Valle del Cauca. Se piensa que quizás la constitución de la familia y los patrones educativos de esta zona del país predispongan a esta situación.

   - Como primera medida un porcentaje alto como el que se menciona no indica si hay "muchas" o "pocas" prostitutas, sólo dice que de las que hay (cuántas?) el 80%

son del Valle del Cauca. Así por ejemplo dicha afirmación se cumpliría, si en Cali sólo hubiera 10 prostitutas y 8 de ellas hubiesen nacido en el Valle.

En segundo lugar, en el supuesto caso de que la prostitución en Cali fuera alta, para atribuir ésta, a una causa específica del Valle, debería compararse dicho índice con el del resto del país.

4. Aunque para la mayoría de la gente los infartos cardíacos están asociados con períodos de ejercicio violento, es mucho más probable que éstos ocurran durante períodos de descanso; más de la mitad de las víctimas de ataques coronarios lo han presentado mientras dormían o descansaban. Menos del 2% lo han presentado mientras estaban dedicados a "hacer deporte, correr o a empujar un gran peso" (tomado de Patterns of Disease, Parke Davis Co.)

   - Observe que los porcentajes a que hace referencia el enunciado se expresan con base en los muertos, por tanto no indican riesgo. Nótese la diferencia entre:

   f(E/M) = porcentaje de los muertos, que hacían ejercicio violento cuando murieron.

   f(M/E) = porcentaje de los que hacen ejercicio violento, que porcentaje muere mientras lo hace.
   En forma análoga debe interpretarse:

   f(D/M) y *f(M/D)* donde la *"D"* hace referencia a "descansar".

   f(M/E) y *f(M/D)* representan el riesgo de morir mientras se hace ejercicio violento o mientras se descansa, respectivamente, valores que aunque tienen más valor para obtener la conclusión mencionada, también deben tratarse con cuidado, pues la edad y la probabilidad de estar haciendo ejercicio violento y la probabilidad de estar descansando en un momento dado son factores importantes, que pueden obrar como factores de confusión.

5. De los registros de accidentes de una secretaría de tránsito, se observó que en el 80% de los accidentes, los involucrados son hombres y sólo en el 20% son mujeres; lo cual demuestra en forma contundente que las mujeres son más cuidadosas que los hombres en la conducción de vehículos automotores.

   - Obsérvese que los porcentajes hacen referencia a los accidentados y no a los conductores en general, ni al tiempo que gastan al volante en un período dado. Por tanto se espera que si son muchas más las horas al volante de los hombres que de las mujeres, haya más accidentes en los cuales haya hombres comprometidos, sin que esto indique un menor cuidado por parte de los hombres.

Sería más adecuado comparar la proporción de accidentes por cada 1.000 horas al volante para mujeres y para hombres.

3.3.2 Independencia estadística entre dos características.

Es imposible referirse a la independencia de dos características, sin pensar en la dependencia. En muchas ocasiones las personas pueden haber notado cierto tipo de asociación entre dos variables, por ejemplo: refiriéndose a las personas "adultas" de cierta ciudad, piénsese en la "dependencia" entre las características: "tener carro propio" y "saber leer", una distribución que seguramente podría aceptarse como ejemplo es la siguiente, con base en una población de 100000 "adultos".

LEE / CARRO PROPIO	SI	NO	
SI	9.900	100	10.000
NO	60.100	29.900	90.000
	70.000	30.000	

Analizando la estructura del cuadro anterior, pueden deducirse los siguientes resultados:

- La población tiene un 30% de personas analfabetas.

- El porcentaje de analfabetas entre los que tienen carro es:

$$f(A/C) = \frac{100}{10.000} = 1\%$$

Donde *A* representa "analfabeta" y *C* representa tener carro.

- El porcentaje de analfabetas entre los que no tienen carro es:

$$f(A/\overline{C}) = \frac{29.900}{90.000} = 33.2\%$$

Con los cálculos realizados puede notarse que la distribución porcentual de los analfabetas es distinta para la subpoblación de los que tienen carro que para los que no tienen carro, es decir:

$$f(A/C) \neq f(A/\overline{C}) \quad \text{y}$$

Lógicamente:

$$f(\overline{A}/C) \neq f(\overline{A}/\overline{C})$$

por tal razón se dice que las características "tener carro propio" y "saber leer" son estadísticamente dependientes.

Se habla de dependencia "estadística" puesto que ésta se concluye sólo con base en la observación de una(s) distribución(es) de frecuencia y no por el análisis cualitativo del fenómeno en el que participan las características en estudio, por eso es importante recalcar que LA DEPENDENCIA ESTADÍSTICA NO EXPRESA RELACIÓN DE CAUSA A EFECTO, aunque pueda usarse como un instrumento preliminar para posteriormente buscar relaciones que permitan dar una explicación al fenómeno en el área específica de estudio.

Definición de independencia estadística entre variables

En resumen se dirá que dos (2) variables *X* e *Y* son **estadísticamente independientes** si la distribución de la característica *X* es la misma en cualquier subconjunto de elementos definidos por la característica *Y*. En forma perfectamente simétrica podrá intercambiarse *X* por *Y*.

Lo anterior puede escribirse con símbolos de varias formas:

X e *Y* son estadísticamente independientes si:

$f(X_i/Y_j) = f(X_i)$ para todo *i, j*

lo cual implica que para cualquier *X* , se cumple:

$f(X_i/Y_1) = f(X_i/Y_2) = \ldots = f(X_i/Y_s) = f(X_i)$

De manera equivalente puede caracterizarse la independencia entre *X* e *Y* por:

$f(Y_j/X_i) = f(Y_j)$ para todo *i, j*.

Por último y recordando la regla de la multiplicación que expresa:

$$f(X_i \cap Y_j) = f(X_i) \cdot f(Y_j/X_i)$$

puede escribirse que:

Las variables *X* e *Y* son estadísticamente independientes si:

$$f(X_i \cap Y_j) = f(X_i)\, f(Y_j)$$

o lo que es igual:

$$f_{ij} = f_{i.}\, f_{.j} \quad \text{, para todo } i, j$$

es decir cuando la distribución conjunta es el producto de las marginales.

Ejemplo 3.8

A continuación se presenta la distribución conjunta de frecuencias absolutas de dos variables *X* e *Y*.

Y / X	Y_1	Y_2	Y_3	Y_4	
X_1	75	90	120	15	300
X_2	100	120	160	20	400
X_3	75	90	120	15	300
	250	300	400	50	1000

La distribución conjunta de frecuencias relativas correspondiente es:

Y / X	Y_1	Y_2	Y_3	Y_4	
X_1	0.075	0.090	0.120	0.015	0.30
X_2	0.100	0.120	0.160	0.020	0.40
X_3	0.075	0.090	0.120	0.015	0.30
	0.25	0.30	0.40	0.05	

Para verificar si las variables *X* e *Y* son estadísticamente independientes de acuerdo con la definición presentada, debe cumplirse para todos los $(X_i \cap Y_j)$ que:

$$f_{ij} = f_{i.}\, f_{.j}$$

Evidentemente si se encuentra alguna pareja (X_i,Y_j) que no satisfaga la definición, es suficiente para concluir que no hay independencia estadística.

Veamos:

$f_{1.}$ x $f_{.1}$ = 0.30 x 0.25 = 0.075 = f_{11}

$f_{1.}$ x $f_{.2}$ = 0.30 x 0.30 = 0.090 = f_{12}

$f_{1.}$ x $f_{.3}$ = 0.30 x 0.40 = 0.120 = f_{13}

$f_{1.}$ x $f_{.4}$ = 0.30 x 0.05 = 0.015 = f_{14}

$f_{2.}$ x $f_{.1}$ = 0.40 x 0.25 = 0.100 = f_{21}

$f_{2.}$ x $f_{.2}$ = 0.40 x 0.30 = 0.120 = f_{22}

$f_{2.}$ x $f_{.3}$ = 0.40 x 0.40 = 0.16 = f_{23}

$f_{2.}$ x $f_{.4}$ = 0.40 x 0.05 = 0.020 = f_{24}

$f_{3.}$ x $f_{.1}$ = 0.30 x 0.25 = 0.075 = f_{31}

$f_{3.}$ x $f_{.2}$ = 0.30 x 0.30 = 0.090 = f_{32}

$f_{3.}$ x $f_{.3}$ = 0.30 x 0.40 = 0.120 = f_{33}

$f_{3.}$ x $f_{.4}$ = 0.30 x 0.05 = 0.015 = f_{34}

Como se verifica la definición para **todo** *i* e *j*, se concluye que las variables *X* e *Y* son estadísticamente independientes.

Estrictamente hablando, esta definición tan rígida, solo se aplica a datos poblacionales y no a datos provenientes de una muestra.
Puede suceder (y es lo más probable) que aun cuando en la población se cumpla en forma exacta la definición, al formar una muestra al azar y aplicar la definición se presentan discrepancias.
El tamaño de estas discrepancias observadas en la muestra permitirán juzgar, con procedimientos de inferencia estadística, que tan plausible (verosímil) es la hipótesis de que en la población se cumple la definición de independencia.

3.3.2.1 Indicadores de dependencia entre variables

Como vimos anteriormente, la dependencia entre dos variables *X* e *Y*, obedece a la definición:

"*X* e *Y* son independientes si y sólo si $f(x_i \cap y_j) = f(x_i).f(y_j)$; para todo *i, j* que es categórica, puesto que no admite término medio: son independientes si cumplió la definición o no son independientes si no cumplió la definición.

En la realidad existen grados o niveles de dependencia que deben ser medidos de manera tal que permita poner en evidencia la intensidad de la dependencia estadística. Puede pensarse en definir un instrumento que involucre la separación entre $f(x_i \cap y_j)$ y el producto $f(x_i) \; . \; f(y_j)$ y que aumente el valor del indicador de dependencia, a medida que se separan los dos términos mencionados.

A continuación se aborda el problema a través de un ejemplo:

Ejemplo 3.9

En la población de "Katherine" se midió la estatura (*Y*) y el peso (*X*) a doscientas personas. Los datos obtenidos se resumen en el cuadro 3.13.

A partir de la distribución conjunta de frecuencias absolutas que muestra el cuadro, se va a tratar de construir algunos indicadores que permitan hacerse idea acerca del grado de dependencia que existe entre las variables peso y estatura para el conjunto de observaciones registradas.

CUADRO 3.13

DISTRIBUCION CONJUNTA DE FRECUENCIAS ABSOLUTAS DE LAS VARIABLES PESO Y ESTATURA.

Y / X	Y_1	Y_2	Y_3	Y_4	Y_5	
X_1	8	11	10	1	0	30
X_2	2	12	14	30	2	60
X_3	10	12	24	17	7	70
X_4	0	5	12	2	21	40
	20	40	60	50	30	

Si aplicamos la definición a dicha distribución, concluimos que no son independientes, puesto que:

$$f(x_1 \cap y_1) = \frac{8}{200} = 0.04$$

$$f(x_1)\, f(y_1) = \frac{30}{200} \times \frac{20}{200} = 0.015$$

de donde se concluye que $f(x_1 \cap y_1) \neq f(x_1).f(y_1)$, pero cuál es el **grado de dependencia** que existe entre *X* e *Y* ?

Para intentar responder esta pregunta, construyamos una distribución conjunta de frecuencias absolutas n^*_{ij} , que satisfaga exactamente la definición de independencia, con el objeto de comparar esta distribución con la distribución real que se tiene; para ello debemos encontrar para cada casilla el valor n^*_{ij} tal que $f^*(x_i \cap y_j) = f(x_i) \,.\, f(y_j)$ Es decir:

$$\frac{n^*_{ij}}{n} = \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}\ , \quad \text{de donde}$$

$$n^*_{ij} = \frac{n_i \times n_{ij}}{n}$$

Así por ejemplo, si *x* e *y* fueran independientes, el valor de la frecuencia absoluta para la casilla correspondiente a $(x_1 \cap y_3)$ debería ser:

$$n^*_{13} = \frac{n_{1.} \times n_{.3}}{n} = \frac{30 \times 60}{200} = 9$$

De esta manera podemos construir la siguiente distribución:

CUADRO 3.14

DISTRIBUCION CONJUNTA DE FRECUENCIAS ABSOLUTAS ESPERADAS, SI LAS VARIABLES PESO Y ESTATURA FUERAN INDEPENDIENTES. (n^*_{ij})

Y / X	Y_1	Y_2	Y_3	Y_4	Y_5	
X_1	3	6	9	7.5	4.5	30
X_2	6	12	18	15	9	60
X_3	7	14	21	17.5	10.5	70
X_4	4	8	12	10	6	40
	20	40	60	50	30	

Del cuadro anterior puede hacerse las siguientes observaciones:

• Algunos valores de las frecuencias absolutas, no son números enteros, por ejemplo el $n^*_{14} = 17.5$, lo cual refuerza la naturaleza hipotética de estos valores.

• Las distribuciones marginales se conservaron en la construcción de la distribución hipotética, es decir:

$$\sum_{j=1}^{s} n^*_{ij} = n_{i.}$$

$$\sum_{i=1}^{m} n^*_{ij} = n_{.j}$$

esto puede deducirse, reemplazando n^*_{ij} por su equivalente $\frac{n_{i.} \times n_{.j}}{n}$ así que:

$$\sum_{j=1}^{s} n_{ij}^{*} = \sum_{j=1}^{s} \frac{n_{i.} \times n_{.j}}{n} = \frac{n_{i.}}{n} \sum_{j=1}^{s} n_{.j} = \frac{n_{i.}}{n} \cdot n = n_{i.}$$

Análogamente para las marginales de *y*.

Para comparar la distribución conjunta de frecuencias absolutas observadas (cuadro 3.13) con la que debería tener si las variables fueran independientes (cuadro 3.14), podríamos calcular las diferencias entre las casillas respectivas y luego hacer la suma, es decir:

$\sum_{i=1}^{m} \sum_{j=1}^{s} \left(n_{ij} - n_{ij}^{*} \right)$, desafortunadamente, esta suma es siempre cero, puesto que:

$$\sum_{i=1}^{m} \sum_{j=1}^{s} \left(n_{ij} - n_{ij}^{*} \right) = \sum_{i=1}^{m} \sum_{j=1}^{s} n_{ij} - \sum_{i=1}^{m} \sum_{j=1}^{s} n_{ij}^{*} = n - n = 0$$

esto nos dice que la suma no puede usarse como indicador del grado de dependencia debido a que su valor es siempre cero, esto puede remediarse, haciendo la suma de los cuadrados de las diferencias, es decir:

$$\sum_{i=1}^{m} \sum_{j=1}^{s} \left(n_{ij} - n_{ij}^{*} \right)^2$$

esta suma será mayor entre mayores sean las diferencias, y será cero sólo cuando todas las casillas coincidan, es decir cuando se cumple la definición de independencia; esto hace que pueda usarse como un indicador de dependencia, pero aún así, presenta algunos inconvenientes como por ejemplo el hecho de dar la misma importancia a diferencias iguales, no importando la magnitud de los valores que se restan, así pues si $n_{ij} = 2$ y $n^{*}_{ij} = 5$ es considerado de la misma manera que si $n_{ij} = 300$ y $n^{*}_{ij} = 303$ y como puede apreciarse aunque en ambos casos hay una diferencia de 3 unidades, ésta es relativamente mayor en el primer caso que en el segundo, de esta manera puede corregirse el indicador expresando la diferencia al cuadrado como una fracción de n^{*}_{ij}, con lo cual resulta el llamado **cuadrado de contingencia**.

$$x^2 = \sum_{i=1}^{m} \sum_{j=1}^{s} \frac{\left(n_{ij} - n_{ij}^{*} \right)^2}{n_{ij}^{*}}$$

el cual puede simplificarse al efectuar el cuadrado y reemplazar a n^{*}_{ij} por su valor $\frac{n_{i.} \times n_{.j}}{n}$, con lo cual se produce la expresión equivalente:

$$x^2 = n \left[\sum_{i=1}^{m} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right]$$

Aunque $x^2 = 0$ sólo cuando las variables son independientes y crece cuando crecen las diferencias, tiene el inconveniente de que está afectado por el número n de observaciones, lo cual no es conveniente, puesto que el grado de dependencia debe medir la diferencia entre $f(x_i \cap y_j)$ y $f(x_i).f(y_j)$ es decir entre

$\frac{n_{ij}}{n}$ y $\frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$ que como puede apreciarse no varía si multiplicamos todas las casillas n_{ij} por una constante k, lo cual es equivalente a multiplicar el número de observaciones por ese mismo factor; este aspecto puede corregirse definiendo el llamado **cuadrado medio de contingencia f^2** .

$$f^2 = \frac{x^2}{n} = \sum_{i=1}^{m}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.} \times n.j} - 1$$

f^2 al igual que x^2 , es siempre mayor o igual que cero y no está acotado en forma general, pero si tiene cota superior para cada problema específico en función del número m de categorías de X y el número s de categorías de la variable Y, esto puede deducirse del hecho:

$$n_{ij} \leq n_{i.}$$
$$n_{ij} \leq n_{.j}$$

de donde resulta que:

$$\frac{n_{ij}^2}{n_{i.} \times n.j} \leq 1$$

se puede demostrar que:

$$0 \leq f^2 \leq min(m-1; s-1)$$

De la anterior expresión se sugiere la construcción de un indicador de dependencia cuyo rango no esté afectado por el número de categorías en X e Y; así surge el llamado **coeficiente de contingencia H^2 de Cramer**

$H^2 = \frac{f^2}{\min(m-1; s-1)}$, con lo cual siempre se garantiza que $0 \leq H^2 \leq 1$

$H^2 = 0$ sólo cuando las variables X e Y son estadísticamente independientes.

$H^2 = 1$ expresa el máximo grado de dependencia, que se presenta cuando a partir del conocimiento de una de las dos características de un elemento, es posible determinar exactamente la característica restante.

En general, a medida que el grado de dependencia aumenta, H^2 se acerca al valor 1.

Para el ejemplo planteado, tenemos:

$n = 200 \quad ; \quad m = 4 \quad ; \quad s = 5$

El cuadrado de contingencia

$$x^2 = \sum_{i=1}^{4}\sum_{j=1}^{5}\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} = \frac{(8-3)^2}{3} + \frac{(11-6)^2}{6} + \frac{(10-9)^2}{9} + \ldots$$

$$\ldots + \frac{(2-10)^2}{10} + \frac{(20-6)^2}{6} = 92.26$$

El cuadrado medio de contingencia

$$f^2 = \frac{x^2}{n} = \frac{92.26}{200} = 0.46$$

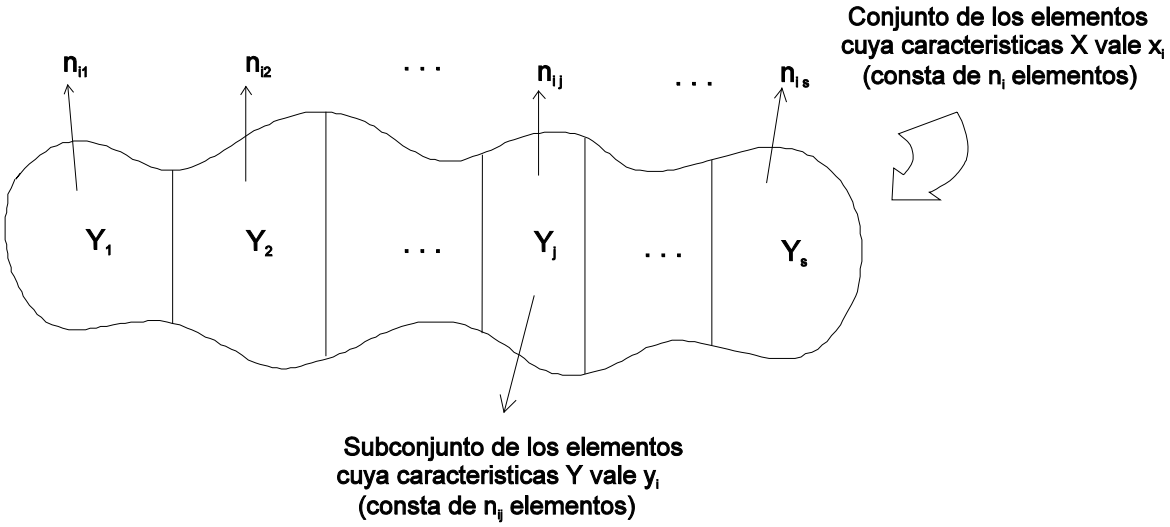
$$H^2 = \frac{f^2}{\min(m-1\,;\,s-1)} = \frac{0.46}{\min(3\,;\,4)} = \frac{0.46}{3} = 0.15$$

Todos los indicadores de dependencia que se han presentado, están definidos bajo el supuesto de que se calculan con base en información poblacional.
En otras palabras, tratan de medir el grado de dependencia de las características sin contemplar el efecto producido por la incertidumbre, cuando se trabaja con una muestra para hacerse una idea sobre la población. No obstante existen pruebas que tienen en cuenta esta incertidumbre.

3.3.3 Media y varianza de distribuciones condicionales

De la misma manera como se presentó el significado de una distribución condicional de frecuencias, considerándola como la distribución de frecuencias de una característica, para un conjunto de elementos que satisfacen cierta condición, puede interesar conocer la media y la varianza para los mencionados elementos, en general podría definirse para ellos cualquier estadígrafo y se estaría refiriendo a **estadígrafos condicionales**, puesto que se calcula para un subconjunto de elementos que satisfacen una condición dada.

Por ejemplo, se podría tener interés en conocer la media aritmética y la varianza de la característica *Y*, para los elementos cuya característica *X* es x_i.



De acuerdo con ésto la media aritmética de *Y* para los que satisfacen la condición $X = x_i$, que denotaremos por $M(Y/x = x_i)$ ó $M(Y/x_i)$, será :

$$M(Y/x_i) = \frac{n_{i1}Y_1 + n_{i2}Y_2 + \ldots + n_{is}Y_s}{n_{i.}}$$

la cual puede escribirse como :

$$M(Y/x_i) = \frac{n_{i1}}{n_{i.}}Y_1 + \frac{n_{i2}}{n_{i.}}Y_2 + \ldots + \frac{n_{is}}{n_{i.}}Y_s$$

Recordando que :

$$f(Y_j/x_i) = \frac{n_{ij}}{n_{i.}}$$

Entonces:

$$M(Y/x_i) = f(y_1/x_i).y_1 + f(y_2/x_i).y_2 + \ldots + f(y_s/x_i).y_s$$

que en representación abreviada es :

$$M(Y/x_i) = \sum_{j=1}^{s} f(Y_j/x_i).y_j$$

De esta manera se pueden calcular tantas medias condicionales, como valores de *x*, así se tendría:

$$M(Y/x_1),\ M(Y/x_2),\ M(Y/x_3),\ldots,\ M(Y/x_m)$$

PROPIEDAD

Un resultado importante es que la media de las medias condicionales coincide con la media de todos los datos.

Si se hace una partición de los elementos de la muestra de acuerdo con los valores de la característica *X*, colocando en un conjunto los que tienen $X = x_1$, en otra los que tienen $X = x_2$ y así sucesivamente, y a cada grupo calculamos la media aritmética, de la característica *Y*, entonces por la propiedad de la media aritmética:

$$\bar{y} = \frac{M(Y/x_1).n_{1.} + M(Y/x_2).n_2 + \ldots + M(Y/x_m).n_{m.}}{n}$$

O lo que es lo mismo:

$$\bar{y} = \sum_{i=1}^{m} M(Y/x_i).f_{i.}$$

En forma perfectamente análoga se podría referir a la media de *X* condicionada por *Y*, $M(X/y_j)$.

Háblese ahora de la varianza de una distribución condicional; así por ejemplo si se quiere calcular la varianza de *Y*, para los elementos que tienen su característica $X = x_i$; se debe recordar que:

$S_y^2 = \frac{1}{n}\sum_{j=1}^{s}\left(y_j - \bar{y}\right)^2.n_{.j}$ Varianza de *Y* para los *n* datos de la muestra.

Si se va a calcular la varianza, sólo para los $n_{i.}$ elementos que satisfacen la condición $X = x_i$ y cuya media aritmética es $M(Y/x_i)$, entonces se escribirá:

$$S_{y/x_i}^2 = \frac{1}{n_{i.}}\sum_{j=1}^{s}\left[y_j - M(Y/x_i)\right]^2.n_{ij}$$

ó

$$S_{y/x_i}^2 = \sum_{j=1}^{s}\left[y_j - M(Y/x_i)\right]^2.\frac{n_{ij}}{n_{i.}}$$

si se tiene en cuenta que

$$f(y_j/x_i) = \frac{n_{ij}}{n_{i.}}$$

Puede escribirse

$$S_{y/x_i}^2 = \sum_{j=1}^{s}\left[y_j - M(Y/x_i)\right]^2.f(y_j/x_i)$$

En forma análoga se puede definir a S^2_{x/y_i}

$$S^2_{x/y_j} = \sum_{i=1}^{m} \left[x_i - M\left(X/y_j\right)\right]^2 . f\left(x_i/y_j\right)$$

Nótese que tanto las distribuciones condicionales de frecuencias como sus rasgos asociados (media condicional y varianza condicional, etc.), no son conceptos nuevos, son exactamente los mismos elementos conocidos, solo que aplicados a un subconjunto de la muestra que satisface una determinada condición. Por lo tanto todas, absolutamente todas las propiedades deducidas para el caso unidimensional se satisfacen en las distribuciones condicionales.

Ejemplo 3.10

Se tomó una muestra de 500 viviendas de la población de Igor y entre otras se observaron las siguientes características: número de personas que duermen en la vivienda (x) y área de dormitorio (Y), en m^2.
Al tratar la información se construyeron las siguientes categorías:

Para la variable *X*
(Número de personas)

X_1: En la vivienda duerme una persona
X_2: En la vivienda duermen dos personas
X_3: En la vivienda duermen tres personas

X_4: En la vivienda duermen cuatro personas
X_5: En la vivienda duermen cinco personas.

Para la variable *Y*
(área de dormitorio en m^2)

Y_1: (3.0, 4.0]
Y_2: (4.0, 6.0]
Y_3: (6.0, 9.0]
Y_4: (9.0, 12.0]
Y_5: (12.0, 16.0]
Y_6: (16.0, 25.0]

De acuerdo con las categorías anteriores se construyó la **distribución conjunta de frecuencias absolutas** para el número de personas y el área de dormitorio como se muestra en el cuadro que aparece a continuación:

Y						

X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	
X_1	10	4	2	5	3	1	25
X_2	4	40	20	15	17	4	100
X_3	3	35	61	10	40	26	175
X_4	1	18	59	14	34	24	150
X_5	2	3	8	6	16	15	50
	20	100	150	50	110	70	500

Estime:

a) El área de dormitorio promedia para las viviendas en que duermen dos personas.

$$M(Y/x_2) = \sum_{j=1}^{6} Y_j' \, f(y_j/x_2)$$

los y_j' son las marcas de clase respectivas

j	Y'_j	$f(y_i / x_2)$
1	3.5	0.04
2	5.0	0.40
3	7.5	0.20
4	10.5	0.15
5	14.0	0.17
6	20.5	0.04

De esta manera:

$M(Y/x_2)$ = 3.5 x 0.04 + 5.0 x 0.40 + ... + 20.5 x 0.04 = 8.415 m^2.

Es decir que las viviendas en que duermen dos personas tienen en promedio un área de dormitorio de 8.415 m^2.

b) La varianza del área de dormitorio, en las viviendas en que duermen dos personas.

$$S^2_{Y/x_2} = \sum_{j=1}^{6} \left[y_j' - M(Y/x_2) \right]^2 \cdot f(y_j/x_2)$$

Como ya se calculó $M(Y/x_2)$ = 8.415

S^2_{Y/x_2} = (3.5 - 8.415)^2 x 0.04 + (5.0 - 8.415)^2 x 0.40 + (7.5 - 8.415)^2 x 0.20 + ... + (20.5 - 8.415)^2 x 0.04

= 17.6 m^4

c) El número de personas promedio que duermen en las viviendas cuya área de dormitorio está entre 4.0 m^2 y 6.0 m^2.

$$M(X/y_2) = \sum_{i=1}^{5} x_i \cdot f(x_i/y_2)$$

i	X_i	$f(x_i / y_2)$
1	1	0.04
2	2	0.40
3	3	0.35
4	4	0.18
5	5	0.03

De esta manera:

$M(X/Y_2)$ = 1 x 0.04 + 2 x 0.40 + 3 x 0.35 + 4 x 0.18 + 5 x 0.03 = 2.76

Es decir que en las viviendas con área de dormitorio entre 4.0 y 6.0 m^2, en promedio duermen 2.76 personas.

d) La varianza del número de personas que duermen en viviendas con área de dormitorio entre 4.0 y 6.0 m^2.

$$S^2_{x/y_2} = \sum_{i=1}^{5} [X_i - M(X/Y_2)]^2 \cdot f(x_i/y_2)$$

Como ya se tiene calculado $M(X/Y_2)$ = 2.76

S^2_{x/y_2} = (1 - 2.76)^2 x 0.04 + (2 - 2.76)^2 x 0.40 + (3 - 2.76)^2 x 0.35 +
(4 -2.76)^2 x 0.18 + (5 - 2.76)^2 x 0.03 = 0.80 (personas)^2

e) El promedio y la varianza del área de dormitorio:

$$\bar{y} = \sum_{i=1}^{6} y'_j \cdot f_{.j}$$

= 3.5 x 0.04 + 5.0 x 0.20 + 7.5 x 0.30 + 10.5 x 0.10 + 14.0 x 0.22 + 20.5 x 0.14
= 10.39 m^2

$$S^2_y = \sum_{i=1}^{6} \left(y'_j - \bar{y}\right)^2 \cdot f_{.j} = 27.4\,\text{m}^2$$

f) El promedio y la varianza del número de personas que duermen por vivienda

$$\overline{X} = \sum_{i=1}^{5} x_i f_{i.}$$

= 1 x 0.05 + 2 x 0.20 + 3 x 0.35 + 4 x 0.30 + 5 x 0.10
= 3.2 personas

$$S_x^2 = \sum_{i=1}^{5} \left(x_i - \bar{x}\right)^2 f_{i.}$$

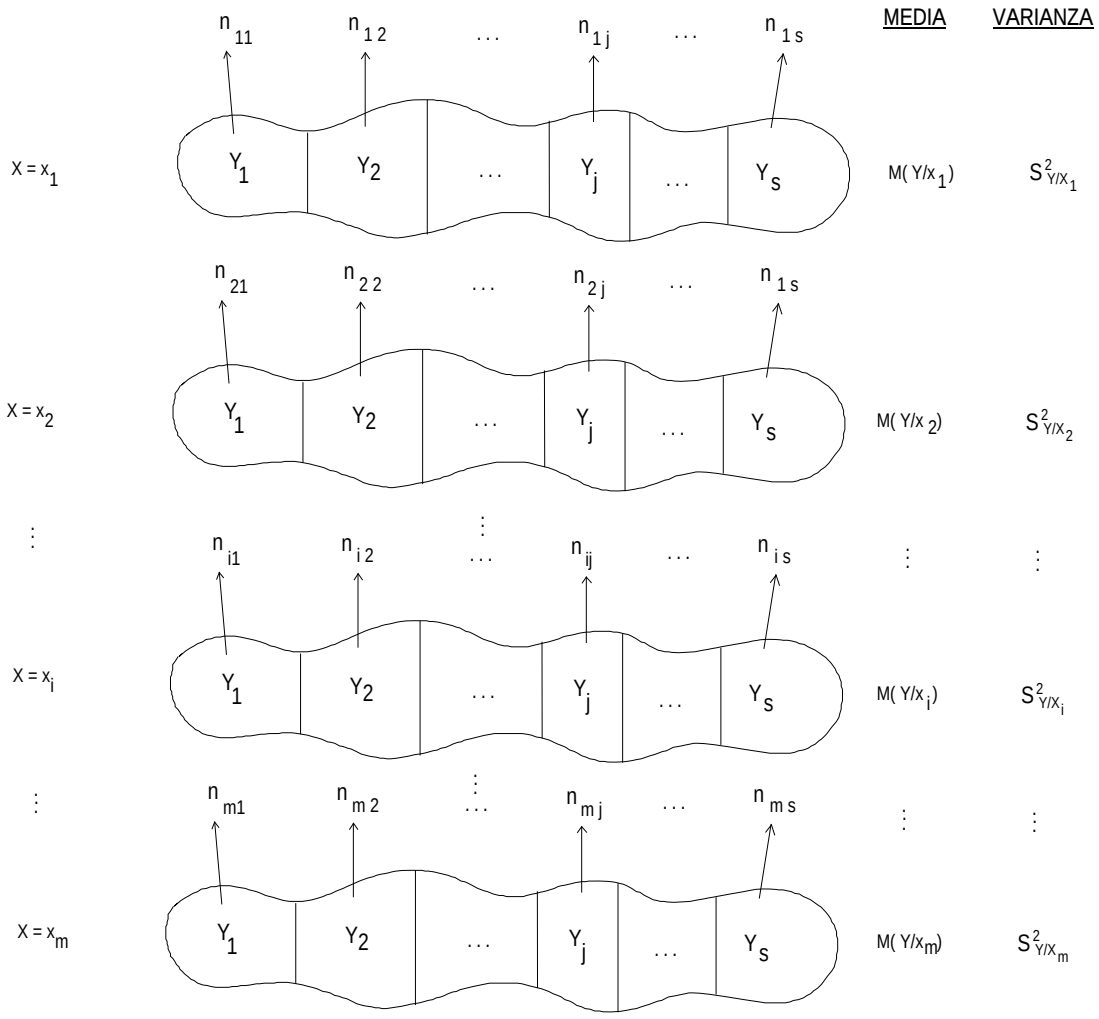
= $(1 - 3.2)^2$ x 0.05 + $(2 - 3.2)^2$ x 0.20 + ... + $(5 - 3.2)^2$ x 0.10
= 1.06 $(\text{personas})^2$

3.3.4 Otra manera de detectar asociación estadística entre características de una población.

En cuanto se trató el concepto de independencia estadística, se enunció que dos características *X* y *Y* son independientes en una población, si la distribución de *Y* es la misma para cualquier subpoblación definida por una condición expresada en términos de la variable *X* . En otras palabras la distribución de *Y* es la misma en todos los subgrupos que se construyan con base en la variable *X* .
A continuación se plantea un interesante procedimiento que compara indirectamente las distintas distribuciones con base en la diferencias entre sus medias aritméticas y escalando estas diferencias al compararlas con la magnitud de las diferencias que pueden ocurrir al interior de un mismo grupo. En resumen compara la variación en grupos (usando la media) con las variaciones internas de los grupos que se comparan. Surgen aquí los conceptos de Intervarianza e Intravarianza.

3.3.4.1 Intervarianza e intravarianza

Si se parte de que la muestra está particionada en subconjuntos de acuerdo con los valores de la característica *X*, la situación sería como muestra el siguiente esquema:



El diagrama muestra que en el subconjunto de elementos que satisfacen $X = x_i$ se puede a su vez clasificar sus elementos de acuerdo con la característica *Y*, y aparece el número de elementos que tendría cada uno de estos nuevos subconjuntos, de acuerdo con la notación establecida.

Cuando se piensa en la variabilidad de la media de *Y*, en los diferentes subconjuntos, es decir cuando se hace referencia a la varianza de $M(Y/x_1)$, $M(Y/x_2)$,..., $M(Y/x_m)$ se está hablando de la **intervarianza**, que se denotará por $S^2_{by(x)}$, de esta manera y teniendo en cuenta que la media aritmética de las medias condicionales es $\bar{y}$ o sea el promedio de *Y* para todos los datos, entonces:

$$S^2_{by} = \sum_{i=1}^{m} \left[M(Y/x_i) - \bar{y}\right]^2 \cdot f_{i.}$$

INTERVARIANZA DE *Y*.
[Varianza de las Medias]

De otro lado cuando interesa formarse una idea sobre la magnitud de la varianza de *Y* al interior de cada subconjunto de datos, es decir cuando se quiere tener una idea sobre la magnitud de las varianzas:
$S^2_{Y/x_1}, S^2_{Y/x_2}, \ \ldots\ , S^2_{Y/x_m}$.

Puede calcularse la **media aritmética** de estas varianzas, la cual se conoce como **intravarianza**, simbolizándola por $S^2_{wy(x)}$
así pues:

INTRAVARIANZA DE *Y*
[Media de las Varianzas]

$$S^2_{wy} = \sum_{i=1}^{m} S^2_{Y/x_i} \cdot f_{i.}$$

Como puede apreciarse la intravarianza no es propiamente una varianza, sino que es un promedio de varianzas.

3.3.4.2 Expresión base del análisis de varianza

Si S^2_y representa la varianza de la característica *Y*, para todos los elementos de la muestra, puede escribirse la expresión:

$$S^2_Y = S^2_{bY(x)} + S^2_{wY(x)}$$

Interesante expresión que representa una versión del conocido **análisis de varianza**, que en palabras diría: la varianza de la distribución marginal de una variable *Y*, se puede siempre expresar como la varianza de las medias condicionadas por alguna característica $X=x$, más la media de las varianzas condicionales por la misma $X=x$.
Antes de probar la expresión base del análisis de la varianza, se presentan algunas observaciones:

1. S^2_y es la varianza de la distribución marginal de la variable *Y*, es decir que no importa si se observaron otras características *X, Z, W*, la varianza de la característica *Y* es S^2_Y , en otras palabras si a los elementos de la muestra no se hubiera observado las características (*X, Y*) sino (*Z, Y*) o (*W,Y*) la varianza de *Y* sería la misma pues se estaría determinando sobre los mismos elementos.

2. $S^2_{bY(x)}$ es la varianza de las medias de *Y* condicionadas por los distintos valores de *X*, que en general depende de la característica condicionante, es decir si las características de interés hubieran sido (*Z,Y*), también podríamos plantear la expresión base del análisis de la varianza:

$$S^2_Y = S^2_{byYz)} + S^2_{wY(z)}$$

Pero en este caso $S^2_{bY(z)}$ representaría la varianza de las medias de *Y*, condicionadas por valores de *Z*; puesto que en general el conjunto de elementos que satis-

facen $X = x$ es distinto al conjunto de elementos que satisfacen $Z = z$, por tanto la varianza de las *M(Y/x)* no tiene que ser igual a la varianza de la *M(Y/z)*; se estaría diciendo con esto que a pesar de que S^2_Y es única para los elementos de la muestra, el valor de la intervarianza S^2_{bY} depende de la característica condicionante de la media de *Y*, lo cual repercute de la siguiente manera:

Dado que la suma de la **intervarianza** S^2_{bY} con la **intravarianza** S^2_{wY} debe dar siempre el mismo valor S^2_y, cualquiera que sea la variable condicionante, entonces si para la variable *X* es mayor el valor S^2_{bY} que para la variable Z, necesariamente la **intravarianza** S^2_{wY} para la condicionante *X*, debe ser menor que para *Z*, de tal manera que la suma siempre arroje el mismo valor S^2_y .

3. Hechas las observaciones anteriores, se discute ahora sobre el significado de la magnitud de la intervarianza $S^2_{bY(x)}$. Si la variable *X* no aporta información para la explicación de la varianza de *Y*, se esperaría que $M(Y/x_i)$ fuera aproximadamente igual para todos los valores de x_i, por ejemplo si se estuviera estudiando las variables ingreso mensual (*Y*) y estatura de la persona (*X*), se espera que el ingreso promedio de las personas con estatura entre 1.60 m y 1.70 m, sea aproximadamente igual al ingreso promedio de las personas con estatura entre 1.70 m y 1.80m y en general para cualquier otro valor de la variable estatura, si se acepta que esta variable no incide en la variación del ingreso mensual. De esta manera se estaría diciendo en el caso planteado, que la varianza de las medias de *Y* condicionadas por *X* (intervarianza) está cerca a cero y en consecuencia la intravarianza $S^2_{bY(x)}$ será aproximadamente igual a $S^2_{Y.}$.
   Análogamente, si la variable *X* influye bastante en la variación de la variable *Y*, se espera que la media condicionada de *Y* sufra "variaciones significativas" cuando se calcula para distintos valores de la condición dada por *X*, por ejemplo si entre las variables de peso (*Y*) y estatura (*X*) existe una fuerte asociación en el sentido de que la estatura explica la variación del peso en un conjunto de personas de una muestra, se espera que haya variaciones en el peso promedio de las personas que tienen entre (1.40,1.50) de estatura y el peso promedio de las que tienen entre (1.50,1.60) y en las que tienen (1.60 y 1.70), etc.; o sea que el valor de $M(Y/x_i)$ depende de la categoría x , que se estudie, esto significa que la varianza de las $M(Y/x_i)$, es decir la **intervarianza**, es "grande".

   Cuando usamos la palabra "grande", lo hacemos en sentido relativo, puesto que siempre se cumple que:

$$0 \leq S^2_{bY} \leq S^2_Y$$

   Entonces S^2_{bY} será más grande, cuanto más cerca esté de $S^2_{Y.}$

   El caso extremo de máxima fuerza de *X* en la explicación estadística de la variación de *Y* se cumpliría, cuando S^2_{bY} tome su máximo valor S^2_Y y en consecuencia $S^2_{wY} = 0$, puesto que la suma de S^2_{wY} y S^2_{bY} siempre da S^2_Y; la

situación planteada anteriormente ocurriría cuando todos los elementos del conjunto de los que satisfacen $X = x_i$, tienen exactamente el mismo valor de *Y*, es decir cuando $S^2_{Y/xi} = 0$ para todo x_i, lo cual indicaría que existe una relación funcional entre *X* e *Y* (esto significa que para un valor dado de *x* existe un único valor de *y*).

Como ayuda nemotécnica de S^2_{bY} y de S^2_{wY}, son del inglés "between" y "within" que significa "entre" y "dentro" respectivamente (en castellano ambas intervarianza e intravarianza tienen las mismas iniciales).

3.3.4.3 Razón de correlación

Ya se había dicho que si la fuerza de *X* en la explicación de la variación de *Y*, es "grande", entonces la intervarianza de *Y* será "grande" comparada con su valor máximo posible, puesto que:

$$0 \leq S^2_{bY} \leq S^2_Y$$

Este hecho permite expresar la intervarianza como fracción de la varianza total S^2_Y, así se define la razón de correlación:

$$e^2_{y.x} = \frac{S^2_{by}}{S^2_y}$$

De esta manera se tiene que:

$$0 \leq e^2_{y.x} \leq 1$$

Si $e^2_{y.x} = 0$, indica que el promedio de *Y* en el subconjunto de elementos que satisfacen $x = x_i$, es la misma para todo x_i, es decir, el factor *X* no tiene incidencia estadística en la variación de la variable *Y*.

Si $e^2_{y.x} = 1$, indica que $S^2_{by} = S^2_y$ y en consecuencia $S^2_{wy} = \sum_{i=1}^{m} f_{i.} S^2_{y/x_i} = 0$, lo cual implica que todas las $S^2_{y/x_i} = 0$, es decir que al interior del conjunto en el cual $X = x_i$, *Y* es una constante, este hecho marca el mayor grado de fuerza de *X* en la variación de *Y*, puesto que el valor de *X* determinaría en forma inequívoca el valor de la característica *Y*.

En general entre mayor sea el valor de $e^2_{y.x}$ más importante será el factor (variable) *X*, en la explicación de la variación de la característica *Y*.

Recuerde que en la notación $e^2_{y.x}$ se quiere indicar que es de interés la variación de *Y*, cuando la variable *X* está condicionando.

Si se escribiera $e^2_{x.y}$, se hace referencia a la variación de *X*, cuando es *Y* la característica condicionante. En general $e^2_{x.y}$ y $e^2_{y.x}$ son distintos.

Ejemplo 3.11

Con base en la información suministrada en el ejemplo 3.10 de la pág. 128 calcule la intravarianza y la intervarianza para la variable "área de dormitorio", condicionada por la variable "número de personas que duermen en la vivienda" y opine sobre la asociación estadística de las mismas.

Como la intervarianza $S^2_{by} = \sum_{i=1}^{m} \left[M(Y/x_i) - \bar{y}\right]^2 \cdot f_{i.}$

y la intravarianza : $S^2_{wy(x)}$

$$S^2_{by} = \sum_{i=1}^{m} S^2_{y/x_i} \cdot f_{i.}$$

Esto significa que se debe calcular previamente $M(Y/x_i)$, f_i, $S^2_{y/xi}$ para cada *i*.

Sabiendo que:

$$M(Y/x_i) = \sum_{j=1}^{m} y'_j \, f(y_j/x_i)$$

$$S^2_{y/x_i} = \sum_{j=1}^{m} \left[y_j - M(Y/x_i)\right]^2 \cdot f(y_j/x_i)$$

Se puede construir el siguiente cuadro con la información

i	x_i	$M(Y/x_i)$	$S^2_{y/xi}$	$f_{i.}$
1	1	7.400	21.02	0.05
2	2	8.415	17.60	0.20
3	3	10.520	27.68	0.35
4	4	11.006	26.01	0.30
5	5	13.53	29.95	0.10

$\bar{y} = 10.39$

La intervarianza

$S^2_{bY(x)}$ = $(7.40 - 10.39)^2$ x 0.05 + $(8.415 - 10.39)^2$ x 0.20 + ...
... + $(13.53 - 10.39)^2$ x 0.10 = 2.33

la intravarianza

$S^2_{wY} = 21.02 \times 0.05 + 17.60 \times 0.20 + ... + 29.95 \times 0.10 = 25.06\ m^4$

$S^2_{bY} + S^2_{wY} = 2.33 + 25.06 = 27.4$ valor que coincide con la varianza S^2_Y calculada en el ejemplo 3.10.

En este caso: $e^2_{y.x} = \frac{S^2_{by}}{S^2_y} = \frac{2.33}{27.4} = 0.085$

Es decir que la intervarianza representa el 8.5% de la variación de *Y*; lo cual significa que la variable "número de personas que duermen en la vivienda" tiene muy poca fuerza en la explicación estadística de la variación de la variable "área de dormitorio". Es decir que cuando *X* varía el promedio de *Y* no varía mucho.

Nótese que la intervarianza está midiendo cuan distintos son los promedios de la variable *Y* cuando se calculan en diferentes conjuntos de acuerdo con la característica *X*, si la intervarianza es pequeña, como este caso, indica que esas medias condicionadas son muy similares no importa en cual conjunto de $X = x_i$, se calcula; aquí se estaría diciendo que el área promedia de dormitorio para las viviendas en que duerme una persona es similar al área promedio para las viviendas en que duermen dos personas, y al área promedio de las viviendas formando la muestra global.

Ejemplo 3.12

Si se repite el ejemplo anterior pero realizando el análisis de la varianza a la variable "número de personas que duermen en la vivienda" (*X*), condicionada por la variable "área de dormitorio" (*Y*).

En este caso las expresiones a calcular son:

La intervarianza

$$S^2_{bx} = \sum_{j=1}^{s} \left[M\left(X/y_j\right) - \bar{x} \right]^2 f_{.j}$$

La intravarianza

$$S^2_{wx} = \sum_{j=1}^{s} S^2_{x/y}\, f_{.j}$$

Para computar la intervarianza y la intravarianza se requiere del cálculo previo de:

$$M\left(X/y_j\right)=\sum_{i=1}^{m} x_i f\left(x_i/y_j\right)$$

Para $j = 1, 2, \ldots, 5$

$$S^2_{x/y_j}=\sum_{i=1}^{m}\left[x_i - M\left(X/y_j\right)\right]^2 \cdot f\left(x_i/y_j\right)$$

Esto significa que se debe contar con la distribución condicional de *X* dado *Y* o de la distribución conjunta, para con base en ella calcularlas, por tanto se escribirá (tomándola del enunciado original del ejemplo 3.10)

CUADRO 3.15

DISTRIBUCION CONJUNTA DE FRECUENCIAS ABSOLUTAS
PARA LAS VARIABLES *X* e *Y*

Y / X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	
X_1	10	4	2	5	3	1	25
X_2	4	40	20	15	17	4	100
X_3	3	35	61	10	40	26	175
X_4	1	18	59	14	34	24	150
X_5	2	3	8	6	16	15	50
	20	100	150	50	110	70	500

Con base en la anterior información se llena el siguiente cuadro:

CUADRO 3.16

j	Intervalo $(L_{j-1}, L_j]$	$f_{j.}$	$M(Y/x_j)$	S^2_{x/y_j}
1	(3.0 , 4.0]	0.04	2.05	1.7475

2	(4.0 , 6.0]	0.20	2.76	0.8024
3	(6.0 , 9.0]	0.30	3.34	0.6777
4	(9.0 , 12.0]	0.10	3.02	1.4596
5	(12.0 , 16.0]	0.22	3.39	1.0079
6	(16.0 , 25.0]	0.14	3.69	0.8125

$$\overline{X} = \sum_{i=1}^{m} f_{i.} x_i = \ 0.05 \text{ x } 1 + 0.20 \text{ x } 2 + \ ... \ + 0.10 \text{ x } 5 \ = \ 3.2$$

Así que :

LA INTERVARIANZA

$S^2_{bx(y)}$ = $(2.05\text{-}3.2)^2$ x 0.04 + $(2.76\text{-}3.2)^2$ x 0.20 + ... + $(3.69\text{-}3.2)^2$ x 0.14
= 0 .1423

LA INTRAVARIANZA

$S^2_{wx(y)}$ = 1.7475 x 0.04 + 0.8024 x 0.20 + ... + 0.8125 x 0.14
= 0.9151

Si se calcula $S_x^2 = \sum (x_i - \bar{x})^2 \cdot f_{i.} = 1.06$

Se puede comprobar de nuevo que:

$$S^2_x = S^2_{bx(y)} + S^2_{wx(y)}$$

Calculando $e^2_{x.y} = \frac{S^2_{bx}}{S^2_x} = 0.132$

indica que la intervarianza de *X* es aproximadamente el 13.2% de la varianza de *X* en la muestra.

Expresión fundamental del análisis de varianza. Una prueba:

$$S^2_y = S^2_{bY(x)} + S^2_{wY(x)}$$

Donde

$$S^2_{by} = \sum_{i=1}^{m} \left[M(Y/x_i) - \bar{y} \right]^2 \cdot f_{i.}$$

$$S^2_{wy} = \sum S^2_{y/x_i} \cdot f_{i.}\,;\; S^2_{y/x_i} = \sum_{j=1}^{s} \left(y_j - M(Y/x_i) \right)^2 f\left(y_j / x_i \right)$$

Se sabe que

$$S^2_y = \sum_{j=1}^{s} \left(y_j - \bar{y} \right)^2 f_{.j} \;, \quad \text{si se tiene en cuenta que}$$

$$f_{.j} = \sum_{i=1}^{m} f_{ij}$$

Se puede escribir a S^2_Y como:

$$S^2_y = \sum_{i=1}^{m} \sum_{j=1}^{s} \left(y_j - \bar{y} \right) f_{ij}$$

Sumando y restando $M(Y/x_i)$ dentro del paréntesis:

$$S^2_y = \sum_{i=1}^{m} \sum_{j=1}^{s} \left\{ y_j - M(Y/x_i) + M(Y/x_i) - \bar{y} \right\}^2 f_{ij}$$

Desarrollando el cuadrado obtenemos:

$$S^2_y = \sum_{i=1}^{m} \sum_{j=1}^{s} \left[y_j - M(Y/x_i) \right]^2 f_{ij} + \sum_{i=1}^{m} \sum_{j=1}^{s} \left[M(Y/x_i) - \bar{y} \right]^2 f_{ij} + \\ + 2 \sum_{i=1}^{m} \sum_{j=1}^{s} \left[y_j - M(Y/x_i) \right] \left[M(Y/x_i) - \bar{y} \right] \cdot f_{ij} \qquad \text{(E 3.1)}$$

Se va ahora a mostrar que el primer término es $S^2_{wy(x)}$, que el segundo término es $S^2_{by(x)}$ y que el tercer término vale cero.

Aplicando el principio de multiplicación, se puede escribir f_{ij} como:

$$f_{ij} = f(y_j/x_i) \cdot f_{i.}$$

de esta manera el primer término queda:

$$\sum_{i=1}^{m}\sum_{j=1}^{s}\left[y_j - M(Y/x_i)\right]^2 f\left(y_j/x_i\right)\cdot f_{i.} =$$

$$\downarrow$$

$$S^2_{y/x_i}$$

Entonces

$$= \sum_{i=1}^{m} S^2_{y/x_i} \cdot f_{i.}$$ expresión ésta que corresponde a la **intravarianza** $S^2_{wy(x)}$.

Véase ahora que el segundo término corresponde a la intervarianza $S^2_{by(x)}$:

$$\sum_{i=1}^{m}\sum_{j=1}^{s}\left[M(Y/x_i)-\bar{y}\right]^2 f_{ij} = \sum_{i=1}^{m}\left[M(Y/x_i)-\bar{y}\right]^2 \sum_{j=1}^{s} f_{ij}$$

$$= \sum_{i=1}^{m}\left[M(Y/x_i)-\bar{y}\right]^2 f_{i.} = S^2_{by}$$

Por último se prueba que el tercer término de la expresión (E 3.1) vale siempre cero:

$$2\sum_{i=1}^{m}\sum_{j=1}^{s}\left[y_j - M(Y/x_i)\right]\left[M(Y/x_i)-\bar{y}\right]f_{ij} =$$

$$2\sum_{i=1}^{m}\sum_{j=1}^{s}\left[y_j - M(Y/x_i)\right]\left[M(Y/x_i)-\bar{y}\right]f\left(y_j/x_i\right)\cdot f_{i.}$$

$$=2\sum_{i=1}^{m}\left[M(Y/x_i)-\bar{y}\right]f_{i.}\cdot\sum_{j=1}^{s}\left[y_j - M(Y/x_i)\right]f\left(y_j/x_i\right)$$

ya que el promedio de las medias condicionales es $\bar{y}$, es decir:

$$\sum_{i=1}^{m}\left[M\left(Y/x_i\right)-\bar{y}\right]\cdot f_{j.}=\sum_{i=1}^{m}M\left(Y/x_i\right)-\bar{Y}=0$$

de esta manera se ha probado que

$$S^2_y \;=\; S^2_{by} \;+\; S^2_{wy}$$

3.3.5 La covarianza y el coeficiente de correlación entre dos variables.

Con los conceptos de independencia estadística, se construyen algunos indicadores de asociación estadística, que se basan esencialmente en la expresión de las diferencias de las distintas distribuciones condicionales (cuadrado de Cramer, f^2,H^2).

Se presenta luego, nuevos elementos de asociación estadística al introducir la expresión fundamental del análisis de la varianza y la razón de correlación, los cuales pretenden plasmar las diferencias entre las distribuciones condicionales, expresada a través de una valoración de la variabilidad de las medias aritméticas condicionales, escalándolas o evaluándolas en comparación con la variabilidad interna de las propias distribuciones condicionales.

Estos instrumentos pretenden detectar asociación estadística en general, es decir no discrimina el sentido de la asociación (su dirección) pero sí dan una idea de la fuerza de la asociación.

A partir de los conceptos de covarianza y correlación lineal, que se desarrollan a continuación, se pretende detectar o conocer sobre la fuerza de asociación estadística de dos variables en la dirección de una línea recta.

Sea (x_1,y_1), (x_2,y_2), ... , (x_n,y_n) una muestra de n elementos a cada uno de los cuales se ha observado las características X e Y.

Si se tuviera interés en calcular la varianza de la variable:

$$T_i \;=\; x_i \;+\; y_i$$

$$S_T^2 = \frac{1}{n}\sum_{i=1}^{n}\left(t_i - \bar{T}\right)^2 \quad \text{es decir que:}$$

$$S_{(x+y)}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i + y_i\right) - \left(\bar{x} + \bar{y}\right)\right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right) + \left(y_i - \bar{y}\right)\right]^2$$

Desarrollando el cuadrado, se obtiene

$$= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 + 2\,\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

o sea que

$$S_{(x+y)}^2 = S_x^2 + S_y^2 + 2\,\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

al término $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$ se le conoce como **covarianza** entre las variables *x* e *y*, que se denotará así:

$$COV(x,y) = \sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n} \qquad \text{(E 3.2)}$$

De esta manera se puede expresar la varianza de *(x + y)* como:

$$S^2_{(x+y)} = S^2_x + S^2_y + 2\,COV\,(x,y)$$

A continuación se explora el significado de la covarianza.

Se observa que si la tendencia es que ambos factores de la expresión (E 3.2) tengan siempre el mismo signo, entonces la covarianza tendría signo positivo. Véase la figura 3.5

Se ha dividido el plano en cuatro cuadrantes:

en el cuadrante I, quedan los puntos para los cuales

$x > \bar{x} \Rightarrow (x - \bar{x}) > 0$
$y > \bar{y} \Rightarrow (y - \bar{y}) > 0$

en el cuadrante II

$x < \bar{x} \Rightarrow (x - \bar{x}) < 0$
$y > \bar{y} \Rightarrow (y - \bar{y}) > 0$

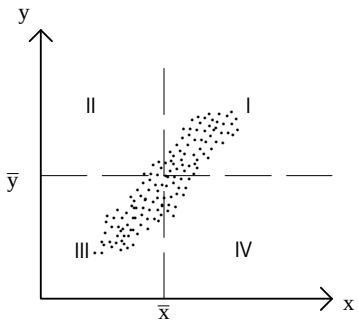


Fig. 3.5

en el cuadrante III

$x < \bar{x} \Rightarrow (x - \bar{x}) < 0$
$y < \bar{y} \Rightarrow (y - \bar{y}) < 0$

y en el cuadrante IV

$x > \bar{x} \Rightarrow (x - \bar{x}) > 0$
$y < \bar{y} \Rightarrow (y - \bar{y}) < 0$

De esta manera si en el diagrama de dispersión los puntos se encuentran con mayor tendencia en los cuadrantes I y III, entonces la covarianza tendrá signo positivo; en cambio si la mayor tendencia está en los cuadrantes II y IV, la covarianza tendrá signo negativo.

Cuando la covarianza es positiva y "grande" indica que hay una **tendencia fuerte** de las variables a crecer en forma conjunta, es decir que cuando *x* crece la tendencia de *y* también es a crecer; lo contrario ocurre cuando la covarianza es negativa y "grande" (en valor absoluto), ver figura 3.6.

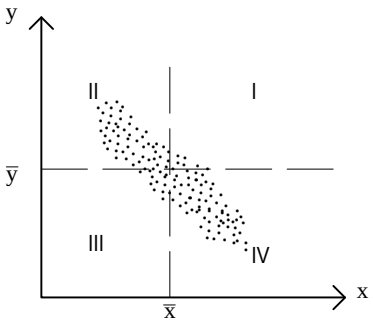


Fig. 3.6

La covarianza proporciona una idea (aunque no muy precisa) sobre el grado de conformación lineal de los puntos en el diagrama de dispersión.

Si el diagrama de dispersión tiene la forma que muestra la figura 3.7

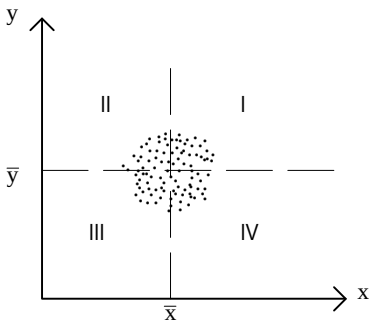


Fig. 3.7

Se observa que para cada punto (x_i, y_i) en un cuadrante existe un simétrico con respecto al eje $x = \bar{x}$ y otro simétrico con respecto al eje $y = \bar{y}$, por tanto $COV(x,y)$ estará próxima a cero.

Antes de seguir concretando los conceptos esbozados, se presentan algunas propiedades de la covarianza.

i) $COV(x + a, y + b) = COV(x,y)$ lo cual significa que la covarianza es invariante con la traslación de los ejes.

ii) $COV(ax, by) = a.b\ COV(x,y)$

De estas propiedades y la definición puede deducirse que

iii) $COV(ax + b, cy + d) = ac\ COV(x,y)$

iv) $COV(x, x) = S^2_x$

Como puede observarse la covarianza es afectada por los cambios de escala, esto hace que su magnitud dependa de las unidades en que se midan las variables x e y, lo cual no es bueno cuando se trata de conocer si la covarianza es "grande" o no, para obtener una idea sobre el grado de relación lineal entre las variables.

Este inconveniente se resuelve al conocer cotas para la covarianza, puesto que:

$$|COV(x, y)| \quad \leq \quad S_x \,.\, S_y$$

Con base en esta propiedad, podremos juzgar si la covarianza entre dos variables es "grande" o "pequeña", comparándola con el producto $S_x \,.\, S_y$.

Mirándolo de otra manera:

$$\frac{\left|COV(x, y)\right|}{S_x \cdot S_y} \leq 1$$

es decir que:

$$-\ 1 \leq \frac{COV(x, y)}{S_x \cdot S_y} \leq\ 1$$

De esta manera si se define el indicador:

$$r = \frac{COV(x, y)}{S_x \cdot S_y}$$

Se sabe que r tiene el mismo signo que la covarianza y además

$$-1 \leq r \leq 1$$

a este indicador se le conoce como **coeficiente de correlación** lineal.

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{S_x \cdot S_y} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)}{S_x} \cdot \frac{\left(y_i - \bar{y}\right)}{S_y}$$

De esta forma si $|r| = 1$ indica que todos los puntos en el diagrama de dispersión tienen una conformación rectilínea perfecta que tendrá pendiente positiva o negativa dependiendo del signo del coeficiente de correlación lineal *r*; por tanto entre más cerca del valor 1 (uno) esté $|r|$, más cercano está el diagrama de dispersión a una conformación rectilínea y entre más cerca a cero esté $|r|$, más lejos estará el diagrama de dispersión a una conformación rectilínea. A continuación se presentan diagramas de dispersión y sus correspondientes coeficientes de correlación lineal.

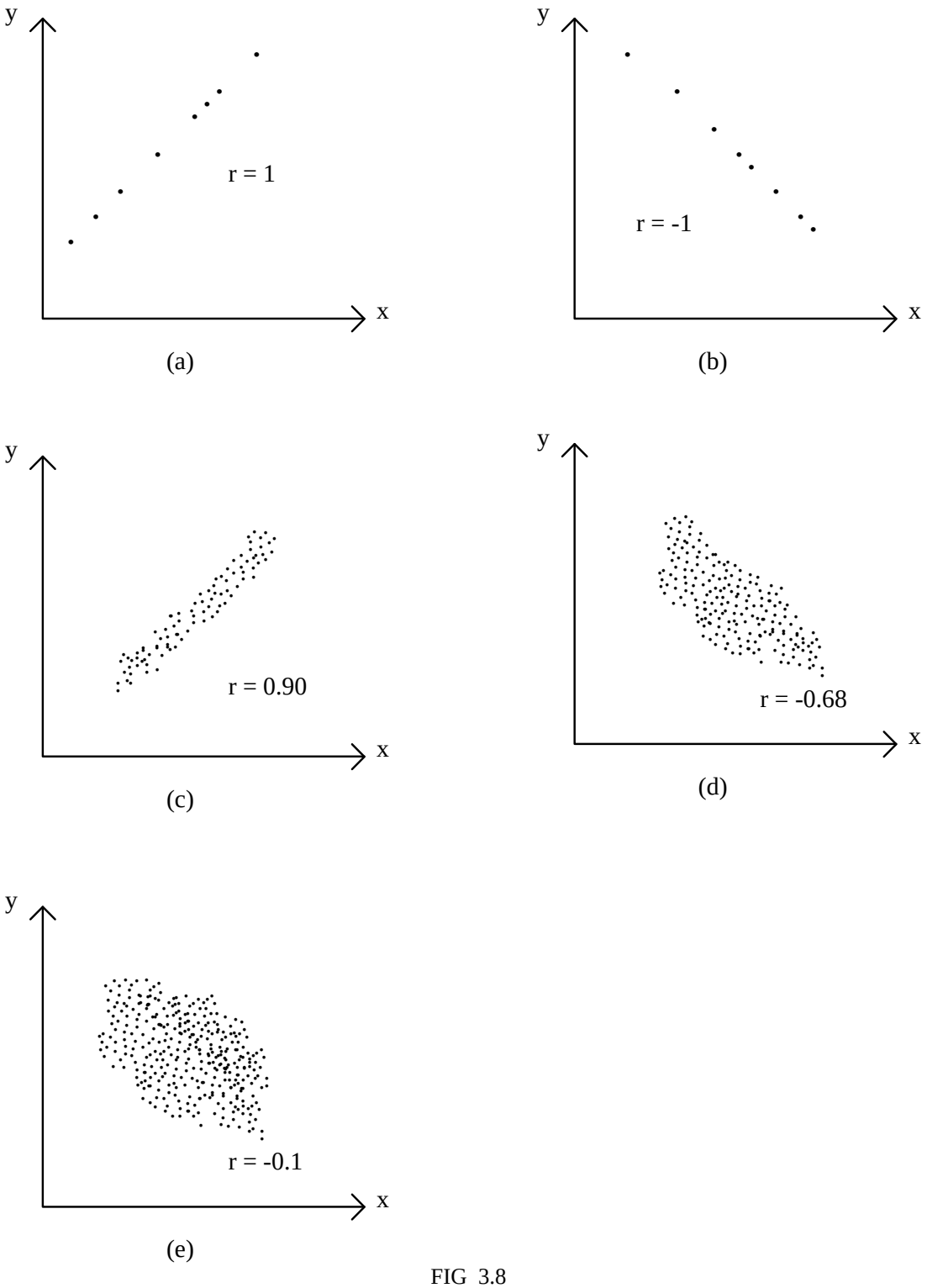


FIG 3.8

Véase ahora, que efectivamente el valor del coeficiente de correlación *r* está siempre en el intervalo [-1, +1]

Sean:

$$Z = \frac{x - \bar{x}}{S_x}$$

$$W = \frac{y - \bar{y}}{S_x}$$

De esta manera $\bar{Z} = 0$ y $S^2_z = 1$ análogamente $\bar{W} = 0$ y $S^2_w = 1$

Se sabe que la varianza de cualquier conjunto de datos es siempre no negativa, por tanto:

a) $V(z + w) \geq 0 \implies V(z) + V(w) + 2\,COV(z, w) \geq 0$

Como $V(z) = 1 = V(w)$, entonces:

$1 + 1 + 2\,COV(z, w) \geq 0$

$2\,[1 + COV(z, w)] \geq 0 \implies \mathbf{COV(z, w) \geq -1}$

b) $V(z - w) \geq 0 \implies V(z) + V(w) - 2\,COV(z,w) \geq 0$

$\implies 2\,[1 - COV(z, w)] \geq 0 \implies \mathbf{COV(z, w) \leq 1}$

De a) y b) se concluye que $-1 \leq COV(z, w) \leq 1$ como

$$COV(z,w) = COV\left(\frac{x - \bar{x}}{S_X}, \frac{y - \bar{y}}{S_X}\right)$$

de acuerdo con las propiedades de la covarianza:

$$COV(z, w) = \frac{1}{S_x \cdot S_y} COV(x, y)$$

$$= \frac{COV(x, y)}{S_x \cdot S_y} = r \quad \text{coeficiente de correlación}$$

por tanto

$$-1 \leq r \leq 1$$

OBSERVACIONES

1. Si los datos están expresados en términos de una distribución conjunta de frecuencias, entonces, la covarianza puede escribirse:

$$COV(x,y)=\sum_{i=1}^{m}\sum_{j=1}^{s}(x_i-\bar{x})(y_j-\bar{y})f_{ij}$$

Si los datos están agrupados en intervalos de clase, entonces los x_i y/o y_j serán las marcas de clase correspondientes.

2. Si *X* e *Y* son variables estadísticamente independientes, entonces:

$$COV(x,y) = 0 \quad \text{y por tanto} \quad r = 0$$

Demostración:

$$COV(x,y)=\sum_{i}^{m}\sum_{j}^{s}(x_i-\bar{x})(y_j-\bar{y})f_{ij}$$

Si *X* e *Y* son independientes entonces :

$$f_{ij} = f_{i.}\ f_{.j}$$

Así que

$$COV(x,y)=\sum_{i}^{m}\sum_{j}^{s}(x_i-\bar{x})(y_j-\bar{y})f_{i.}f_{.j}$$

$$=\sum_{i}^{m}(x_i-\bar{x})f_{i.}\sum_{j=1}^{s}(y_j-\bar{y})f_{.j}$$

$$=\left(\sum_{i=1}^{m}x_i f_{i.}-\bar{x}\right)\cdot\left(\sum_{j=1}^{s}y_j f_{.j}-\bar{y}\right)$$

$$=0\cdot 0$$

$$COV(x,y)=0$$

Como

$$r_{xy}=\frac{COV(x,y)}{S_x\cdot S_y}=\frac{0}{S_x\cdot S_y}=0$$ lo que queda demostrado.

Es decir que si dos variables son estadísticamente independientes entonces están no correlacionadas linealmente, pero **no correlación lineal no implica independencia**, es decir si $r_{yx} = 0$ no implica que X y Y son estadísticamente independientes. Un ejemplo que ilustra esta situación aparece a continuación.

Ejemplo 3.13

La siguiente es la distribución conjunta de frecuencias absolutas de las variables ingresos (Y) y edad (X) para una muestra de 100 personas.

		INGRESO (en miles de pesos)			
	Y X	(125 , 175)	(175 , 225)	(225 , 275)	
E	(15 , 25)	5	10	0	15
D	(25 , 35)	15	15	10	40
A	(35 , 45)	10	16	4	30
D	(45 , 55)	5	9	1	15
		35	50	15	100

$$COV(x,y) = \sum_{i=1}^{m}\sum_{j=1}^{s}(x_i - \bar{x})(y_i - \bar{y})f_{ij}$$

$$= \sum_{i}^{m}\sum_{j}^{s} x_i y_j f_{ij} - \bar{x}\cdot\bar{y}$$

$$\bar{x} = 34.5; \bar{y} = 19.000$$

$$COV(x,y) = 655.500 - (34.5)(19.000)$$

$$= 655.500 - 655.500 = 0$$

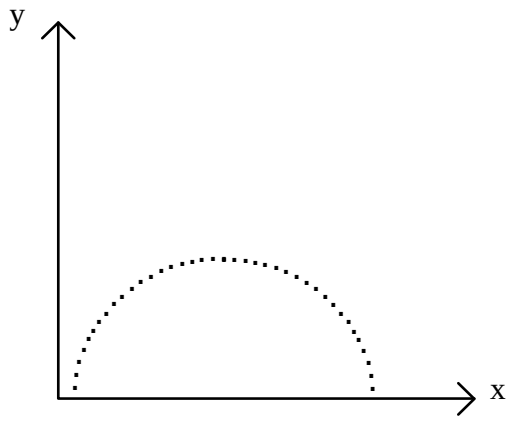
$$r_{xy} = \frac{COV(x,y)}{S_x \cdot S_y} = \frac{0}{S_x S_y} = 0$$

X e Y están no correlacionadas sin embargo, no son independientes, puesto que no cumple que $f_{ij} = f_{i.}\, f_{.j}$ para todo i, j, por ejemplo:

$f_{12} = 0.10 \quad ; \quad f_{1.} = 0.15 \quad ; \quad f_{.2} = 0.5$

de donde se deduce que $f_{12} \neq f_{1.} \cdot f_{.2}$

3. Obsérvese también que en el gráfico, no obstante que el diagrama de dispersión muestra una conformación en la cual aparecen *X* e *Y* conectadas por una relación funcional, sin embargo el coeficiente de correlación lineal es *r = 0*, lo cual indica ausencia de correlación **lineal** y no significa que no exista entre *X* e *Y* otro tipo de correlación.



EJERCICIOS PROPUESTOS

1. A continuación se presenta información de la observación de las variables: número de personas por familia (*X*), e ingreso familiar mensual (*Y*), en 50 familias de la población "Karina"

Familia No.	Número de personas	Ingreso familiar (miles de pesos)
1	4	5110
2	2	4600
3	1	3050
4	4	3920
5	3	3510
6	2	3170
7	2	3860
8	4	2450
9	3	2120
10	3	2040
11	4	2050
12	2	2350
13	2	1980
14	4	1520

Familia No.	Número de personas	Ingreso familiar (miles de pesos)
15	4	1120
16	2	1850
17	2	1980
18	4	1370
19	3	1790
20	2	1540
21	1	910
22	2	810
23	2	1190
24	2	1320
25	3	810
26	4	830
27	4	1770
28	3	1010

Familia No.	Número de personas	Ingreso familiar (miles de pesos)
29	3	1120
30	1	1500
31	1	1100
32	4	920
33	4	1210
34	4	870
35	4	1190
36	1	1560
37	2	840
38	2	960
39	4	810

Familia No.	Número de personas	Ingreso familiar (miles de pesos)
40	2	850
41	2	930
42	2	1000
43	4	850
44	2	1190
45	2	1150
46	4	1690
47	2	1010
48	4	1100
49	2	1180
50	2	1190

1.1 Construya la **distribución conjunta de frecuencias absolutas**, con base en los siguientes intervalos para el ingreso familiar, *Y*, en miles de pesos:

Y_1 : (800;1200] ; Y_2 : (1200;1800] ; Y_3 : (1800;2500] ; Y_4 : (2500;4000]
Y_5 : (4000;6500]

1.2 Construya la función empírica de densidad conjunta.

1.3 Construya la distribución conjunta de frecuencias acumuladas.

1.4 Construya las distribuciones marginales para *X* y para *Y*.

1.5 Construya la distribución condicional del ingreso familiar para las familias con 2 personas.

1.6 Calcule e interprete claramente, de acuerdo con las variables que considera el problema:

$f(x_2/y_3)$, $f(y_3/x_2)$, $f(x_2 \cap y_3)$

$f_{.2}$; $f_{3.}$; $F(2; \$1'500.000)$

1.7 Estime el porcentaje de familias que tienen 2 ó 3 personas y tienen ingresos entre $1'500.000 y $2'700.000.

1.8 Entre las familias que tienen 2 ó 3 personas, qué porcentaje tienen ingresos entre $1'500.000 y $2'700.000.

1.9 Entre las familias que tienen ingresos entre $1'500.000 y $2'700.000, qué porcentaje constan de 2 ó 3 personas.

Calcule:

1.10 El ingreso promedio por familia y su desviación estándar.

1.11 El ingreso promedio por familia, para las familias con 2 personas, y su desviación estándar.

1.12 El número promedio de personas por familia y su desviación estándar.

1.13 El número promedio de personas por familia, para las familias con ingresos entre $2'500.000 y $4'000.000 y su desviación estándar.

1.14 En cuál grupo de familias hay relativamente mayor homogeneidad en el ingreso.

1.15 Son independientes estadísticamente las variables: número de personas por familia y su ingreso. Justifique.

1.16 Si la información en el ejercicio fuera poblacional, cual es el grado de dependencia de las variables.

1.17 Compruebe para la variable ingreso familiar la expresión fundamental del análisis de la varianza:

$$S^2_y = S^2_{by(x)} + S^2_{wy(x)}$$

compare la magnitud de las dos componentes de la varianza, calcule la razón de correlación y comente.

1.18 Con base en diagrama de cajas y alambres, compare la distribución del ingreso para las subpoblaciones definidas por el número de personas en la familia.

2. Se están estudiando las variables continuas X e Y a los elementos de cierta población, en la cual el rango de la variable X es el intervalo (0,1) y el rango de la variable Y es el intervalo (0,4). Si la función de densidad conjunta $f^*(x,y)$, puede expresarse por la función analítica.

$$f^*(x,y) = \begin{cases} axy & \text{si } (x,y) \in D \\ 0 & \text{en cualquier otra parte} \end{cases}$$

Donde D : { $(x,y) / x \in (0,1)$; y $\in (0,4)$ }

2.1 Determine el valor de la constante *"a"*

2.2 Calcule el porcentaje de elementos que tienen $0.2 \le x \le 0.3$ y $2.5 \le y \le 3.8$.

2.3 Entre los que tienen $2.5 \leq y \leq 3.8$, que porcentaje representan los que tienen $0.2 \leq X \leq 0.3$.

2.4 Calcule el porcentaje de los elementos que tienen $0.2 \leq X \leq 0.3$.

2.5 Encuentre la función de distribución acumulativa *F(x;y)*

3. Suponga que las variables *X* e *Y* que se observaron en una población son discretas y sus rangos son respectivamente

$$R_X = \{ 0,1,2 \} \quad : \quad R_Y = \{ 2,3,4 \}$$

Construya una distribución conjunta de frecuencias absolutas de tal manera que el coeficiente H^2 de Cramer valga 1.

4. A continuación se presenta la distribución conjunta de frecuencias absolutas de las variables peso (kg), *X*, y estatura (cms), *Y*, para una muestra de 200 personas adultas observadas en la población de Karen.

Distribución conjunta de frecuencias absolutas para las variables peso (*X*) y estatura (*Y*).

Y / X	Y_1	Y_2	Y_3	Y_4
X_1	5	20	8	7
X_2	12	38	30	20
X_3	3	12	32	13

Donde :

X_1 : (45;55] Y_1 : (150;160]
X_2 : (55;70] Y_2 : (160;165]
X_3 : (70;85] Y_3 : (165;175]
Y_4 : (175;190]

4.1 Construya la función empírica de densidad conjunta para (*X*,*Y*).

4.2 Construya la función empírica de frecuencias acumuladas.

4.3 Construya la función empírica de densidad marginal para la variable peso(*X*).

4.4 Construya la función empírica de densidad del peso, para las personas con estatura entre 165 y 175 cm.

4.5 Construya la función empírica de densidad de la estatura para las personas con peso entre 50 y 60 kg.

4.6 Qué porcentaje de las personas tienen estatura entre 162 y 170 cms. y peso entre 48 y 75 kg.

4.7 De las personas que tienen estatura entre 162 y 170 cms., qué porcentaje de ellas tienen peso entre 48 y 75 kg.

4.8 De las personas que tienen peso entre 48 y 75 kg., qué porcentaje tienen estatura entre 162 y 170 cm.

4.9 Si $F(60, Y_0) = 0.20$; cuál es el valor de Y_0.

4.10 Estime la mediana del "peso".

4.11 Estime la moda de la "estatura"

4.12 Estime el peso promedio y su varianza para las personas con estatura 160 y 175 cm.

4.13 Estime el peso promedio y su varianza para las personas con peso entre 48 y 75 kg.

4.14 Estime el porcentaje de personas para las cuales: su estatura es menor que 2.5 veces su peso.

4.15 Descomponga la varianza de la estatura, con base en los grupos definidos por la variable peso, de acuerdo con la expresión del análisis de la varianza. Comente.

4.16 Compare las distribuciones de la estatura para las subpoblaciones definidas por el peso *X*, con base en diagramas de caja.

5. Muestre que:

$$COV(X,Y) = \frac{\sum X_i Y_i}{n} - \overline{X} \cdot \overline{Y}$$

6. Muestre que el cuadrado medio de contingencia f^2, satisface que:

$$0 \leq f^2 \leq min(m-1 , s-1)$$

donde *m, s,* son el número de categorías de *X* e *Y* respectivamente.

7. En 100 parcelas de igual área, se quiere ensayar tres tipos de abono X_1, X_2, X_3, para evaluar su incidencia en el rendimiento del trigo; para ello se abonaron unas parcelas con X_1, otras con X_2 y otras con X_3 fueron tomadas al azar.
Posteriormente se observó en cada una la producción de trigo Y (en toneladas), registrándose la distribución que aparece a continuación, donde:

Y_1 : (1.0; 1.5] ; Y_2 : (1.5; 2.5] ; Y_3 : (2.5; 3.5] ; Y_4 : (3.5; 5.0]

Distribución conjunta de frecuencias absolutas del rendimiento (*Y*) y tipo de abono (*X*).

Y / X	Y_1	Y_2	Y_3	Y_4	
X_1	7	15	3	5	30
X_2	3	7	10	20	40
X_3	15	8	4	3	30
	25	30	17	28	100

Calcule:

a) $M(Y/X_1)$, $M(Y/X_2)$, $M(Y/X_3)$, $\overline{Y}$

b) S^2_{Y/x_1} , S^2_{Y/x_2} , S^2_{Y/x_3} , S^2_Y

c) Para cuál tipo de abono hay mayor dispersión relativa

d) Haciendo uso de la expresión fundamental del análisis de la varianza, presente un informe sobre incidencia del tipo de abono en el rendimiento del trigo.

8. Con base en los datos del ejemplo 3.13, indique si la variable "edad" explica estadísticamente la variación en la variable "ingreso".

9. Muestre que si X e Y son estadísticamente independientes, entonces la razón de correlación:

$$e^2_{y.x} = e^2_{x.y} =$$

EL MODELO DE REGRESION

ORIGEN[1] DE LA PALABRA "REGRESION" .

Sir Francis Galton fué la primera persona en trabajar con estadística en lo que se refiere a relaciones. A finales del siglo pasado, Galton condujo muchas investigaciones concernientes con la influencia de la herencia sobre varios atributos humanos tanto mentales como físicos. En varios de estos estudios involucró la relación padre-hijo. En particular, Galton (1889) reportó hallazgos acerca de las relaciones entre las estaturas de los padres e hijos. El observo que los padres altos tienden a tener hijos altos y padres bajos tiendes a tener hijos bajos. Sin embargo él también observó lo que llamo efecto de regresión en ésta relación. El notó por ejemplo que la estatura de los hijos tienden a "regresar" a la media de su grupo. Padres muy altos tienden a tener hijos mas altos, pero no tan altos como el promedio de sus padres. Padres de muy baja estatura tienden a tener hijos de baja estatura, pero no tan bajos como el promedio de sus padres. Para aquellos padres en el rango medio, los promedios de las estaturas de sus hijos corresponden mas estrechamente al promedio de la estatura de sus padres.

De esta manera, conociendo la estatura del padre, podría predecirse razonablemente bien, la estatura de su hijo y viceversa. Galton[2] se refirió a este fenómeno como "regresión filial".

El denotó la relación entre la estatura de padres e hijos por la letra "r" (por regresión).

Los términos "línea de regresión" y "ecuación de regresión" corresponden al interés del trabajo específico de Galton. En la actualidad se refieren a una función que es empleada para la "predicción" estadística. Luego la ecuación puede ser referida como "ecuación de predicción".

4.1 INTRODUCCIÓN

En algunas ocasiones es de interés explorar el nivel de asociación estadística entre las mediciones X e Y de dos rasgos de elementos de una población de estudio, con el propósito de usar la información que proporciona una de ellas para tratar de conocer

[1] Lindeman (1980): Introduction to bivariate and multivariate analysis

[2] Sir FRANCIS GALTON. Antropólogo Británico nació en 1822 y murió en 1911. Además de sus invaluables aportes a la teoría de la Herencia y a la estadística, fue quien diseño el sistema de identificación de los individuos humanos con base en la irrepetibilidad de las huellas digitales. (Tomado de 12000 MINIBIOGRAFIAS. Edit. América)

en forma aproximada información sobre rasgos de distribución de la otra característica en un subconjunto dado de elementos en una población. El beneficio que se deriva de llevar a cabo un procedimiento como el planteado es de diversos órdenes, por ejemplo, puede ser más económico observar (medir) la característica *X*, que la característica *Y*, por tal razón sería muy conveniente poder "predecir" rasgos de la distribución de *Y* con base en la observación *X*.

El conocimiento de la relación estadística entre *X* e *Y*, puede traducirse en un ahorro de tiempo, como es el caso de ciertos ensayos en ingeniería tales como el curado del concreto, cuya resistencia máxima se logra a los 28 días; en esta situación es de mucha utilidad disponer de alguna característica que pudiera ser medida más rápidamente y que la asociación de ésta con la resistencia a los 28 días, permitan su estimación. Situaciones como ésta son muy abundantes en las ciencias básicas y también en las acciones de gestión en las cuales la planeación es una etapa fundamental.

Otro tipo de casos en los cuales, cobra importancia el proceso de estimación de una característica con base en otra, es cuando de ordinario, no es posible desde el punto de vista técnico o práctico, la medición directa de la característica *Y*, pero se tienen registros (X_i,Y_i) de algunas ocasiones.

En ocasiones se usa el modelo de regresión, como un instrumento para valorar el impacto de una variable o conjunto de variables en la explicación de la variabilidad de una característica de interés.

En otras oportunidades el interés en la construcción de un modelo de regresión se centra en la estimación e interpretación de algunos de sus parámetros. Casos como estos ocurren por ejemplo en problemas de crecimiento en Biología, o en estimación de coeficientes de elasticidad en Economía. En estas situaciones los esfuerzos no están orientados hacia la predicción.

El modelo de regresión puede ser útil también para detectar la existencia de interacción en el impacto que tienen 2 variables sobre una tercera. Es decir si la magnitud de el efecto de una de ellas depende del valor que asuma la otra característica.

4.1.1 ¿Cuando utilizar un modelo de regresión ?

Son muchas las motivaciones para usar el análisis de regresión, entre las cuales se presentan algunas que no son excluyentes entre sí:

Aplicación # 1.

Se desea caracterizar la relación entre las variables independientes y la dependiente para determinar el grado de dirección y fuerza de asociación. Por ejemplo: se desea medir la fuerza de asociación de las variables: calificaciones del bachillerato, puntajes en el examen del estado (ICFES), tipo de colegio, tiempo transcurrido sin estudiar desde que se graduó de bachiller, edad, sobre la variable dependiente: rendimiento académico en la Universidad del Valle. El objetivo es conocer la importancia relativa de algunos criterios propuestos para el sistema de admisiones.

Aplicación # 2.

Se desea encontrar una fórmula cuantitativa o ecuación para describir (por ejemplo predecir) una variable dependiente *Y* como una función de variables independientes $X_1, X_2, ..., X_n$. La estructura de una cartera en términos del monto por tiempo de atraso influye en el valor mensual del recaudo (*Y*). Se desea predecir el recaudo que se logrará de una cartera con $ X_0 de clientes al día, $ X_1 de clientes con un mes de atraso, $ X_2 con 2 meses de atraso, $ X_3 con 3 meses de atraso, $ X_4 con cuatro (4) o mas meses de atraso.

Aplicación # 3.

Se desea describir cuantitativamente y cualitativamente la relación entre $X_1, X_2, ..., X_k$ y la variable dependiente Y, pero controlando el efecto de otras variables $W_1, W_2, ..., W_p$ que no son propiamente de interés pero que se relacionan con *Y* (estas variables son llamadas factores de confusión o covariables). Ejemplo 1: en un estudio epidemiológico de enfermedades crónicas puede interesar la relación entre la presión sanguínea (Y) y el hábito de fumar (X_1), la clase social (X_2). Se desea controlar la edad (W_1), y el peso corporal (W_2).
Ejemplo 2: se quiere describir la relación entre el conocimiento sobre la regresión lineal (Y) y el método de enseñanza (X_1), controlando el coeficiente de inteligencia (W_1), y estrato social (W_2).

Aplicación # 4.

Se desea saber, entre las variables independientes cuáles son importantes y cuáles no para describir o predecir una variable dependiente. Puede necesitar controlar otras variables.
Ejemplo: una empresa que vende a crédito, desea conocer cuales variables son importantes para el establecimiento del monto a aprobar de un crédito (*Y*). Las variables a considerar son ingreso mensual (X_1), profesión u oficio (X_2), antigüedad en el actual empleo (X_3), vivienda propia (X_4), cuenta bancaria (X_5), barrio de residencia (X_6), número de personas a su cargo (X_7). El estudio se realiza con base a una muestra aleatoria de 1000 clientes, a los cuales se les mide un indicador de cumplimiento (factor de amplificación del plazo), el cual se toma como variable de respuesta.

Aplicación # 5.

Se desea determinar la forma como se relaciona una o más variables independientes con una dependiente *Y*. Aquí el interés está en conocer la estructura del modelo que mejor se ajusta a un conjunto de datos. Al final se sabrá si la relación es rectilínea ó cuadrática ó exponencial ó potencial ó logística, etc.
Ejemplo: se desea conocer la forma de un modelo que relacione la longitud de una especie marina y su edad.

Aplicación # 6.

Se desea comparar la relación entre una(s) variables independientes y otra dependiente (*Y*) en dos o más poblaciones.
Ejemplo 1: determinar si el efecto de fumar (X_1) sobre la presión sanguínea (*Y*), es el mismo en los hombres que en las mujeres, controlando la variable edad (W_1).
Ejemplo 2: comparar si la relación entre el puntaje del examen de admisión (X_1) y el rendimiento en la universidad (*Y*) es la misma para los egresados de los colegios públicos y privados, controlando la variable sexo (W_1).

Aplicación # 7.

Se desea evaluar el "efecto interactivo" de dos o más variables independientes sobre la variable dependiente (*Y*).
Ejemplo 1: se desea determinar si la relación entre el consumo de alcohol (X_1) y la presión sanguínea (*Y*) es diferente dependiendo del consumo de cigarrillos (X_2). la relación entre presión sanguínea y consumo de alcohol puede ser mas fuerte para fumadores empedernidos que para no fumadores. Si esto es cierto, cualquier conclusión sobre la presión y consumo de alcohol, debe tener en cuenta el consumo de cigarrillos.
En general si X_1 y X_2 interactúan en su efecto conjunto sobre *Y*, entonces la relación en *Y* y X_1 depende de los niveles de la otra variable X_2.

Aplicación # 8.

Se desea obtener una estimación válida y precisa de uno ó mas coeficientes de regresión.
Ejemplo 1: coeficiente de elasticidad en el modelo de cantidad vendida y precio.
Ejemplo 2: en un modelo de crecimiento de peces (o de bosques) uno de los parámetros (K) representa la tasa media de crecimiento, su estimación constituye el objetivo central del ajuste de un modelo de regresión.

En el presente capítulo se trata de desarrollar algunos conceptos que concluyen con la definición de instrumentos que permiten construcción de un modelo, presentando

también una herramienta que permite calificar la bondad del modelo; igualmente se destacarán las limitaciones en la aplicación de los instrumentos que se definen.

Se ha puesto de presente que se va a usar una sola característica en el proceso de predicción de otra, este procedimiento puede generalizarse, de tal manera que pueda involucrarse varias variables como base para la predicción.

4.2 LA LINEA DE REGRESION PROPIAMENTE DICHA

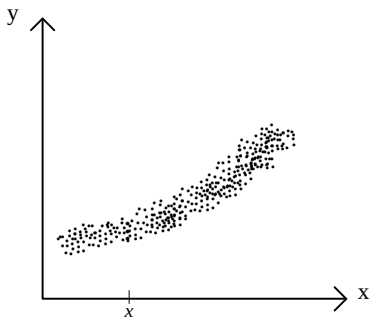


Fig. 4.1

Hay que destacar que en general *Y* no guarda relación funcional con *X*, es decir, existen elementos que teniendo la misma medida en su característica *X*, poseen diferentes valores en la medida de su característica *Y*, como lo muestra la figura 4.1; por ejemplo, dos personas que tengan igual peso corporal, no necesariamente tendrán la misma estatura, puesto que no existe una relación funcional del peso a la estatura; sin embargo el peso de una persona es una información que puede mejorar la "predicción" o "estimación" de su estatura. Si lo miramos un poco intuitivamente, es equivalente a comparar cual estimación se espera sea mejor, cuando se pide "predecir" la estatura que tiene una persona que va a ser extraída al azar de la población *A* o cuando se pide predecir la estatura de una persona que va a ser extraída al azar entre las personas que pesan 70 kg. en la población *A*.

En el peor de los casos se podría decir que el peso no ayuda en la predicción de la estatura y quedaríamos como en la primera situación planteada.

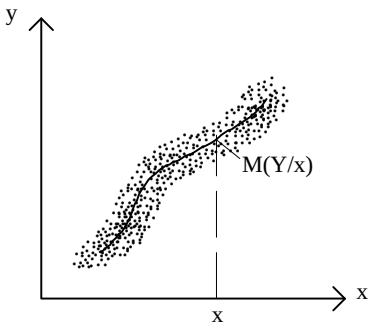


Fig. 4.2

En otras palabras podría decirse que la información sobre el peso de las personas ayuda a mejorar la "predicción" de su estatura, si la varianza de la estatura entre los individuos con el mismo peso corporal es menor que la varianza de la estatura considerando todos los elementos de la población, o sea que más importante será el peso para "predecir" la estatura entre menor sea la **intravarianza** de la estatura comparada con su varianza considerando todos los elementos, es decir, que entre mayor sea la razón de correlación pertinente y en este caso podría usarse para hacer la predicción, la estatura promedia de las personas que pesan 70 kg. para el ejemplo propuesto, y en general podría proponerse *M(Y/x)*, para

predecir la característica *Y* de un elemento que tiene una medida de *x* de su característica *X*, lo cual da origen a la curva que muestra la figura 4.2 en la cual se encuentran los promedios de la variable *Y* para los distintos valores de la variable *X*, a ésta curva se le conoce como línea de regresión propiamente dicha y en este caso se dice que es una línea de regresión de *Y* sobre *X*, para precisar que *Y* es la variable de respuesta que se desea predecir a partir del conocimiento de *X*; esto sugiere que existen dos líneas de regresión una de *Y* sobre *X* y otra de *X* sobre *Y* cuando se requiere *X* como variable de respuesta.

En general estas dos líneas no son coincidentes.

En adelante consideramos la línea de regresión de *Y* sobre *X*, a no ser que se haga explícito lo contrario.

4.3 LA LINEA DE REGRESION MINIMO-CUADRATICA

Idealmente, la línea de regresión que aparece en la figura 4.2 se construiría uniendo a mano alzada las medias condicionales que permita calcular la muestra obtenida, esto significa que si necesita predecir *Y* a partir de un valor *x*, se debería hacer usando el gráfico, puesto que no se tiene un modelo matemático que permita escribir *M(Y/x)* como una función de *x*.

Esta desventaja puede eliminarse si se plantea una familia de modelos y se encuentra, de acuerdo con algún criterio, el modelo de esa familia que "mejor" se ajusta al diagrama de dispersión, como una aproximación a la línea de regresión propiamente dicha.

Cuando se habla de "familia de modelos" en el contexto anterior, se hace referencia por ejemplo a la familia de los modelos rectilíneos, o la familia de modelos parabólicos, familia de polinomios de grado 5, o en general a la familia de modelos que satisfacen una expresión dada.

La determinación de la familia de modelos que se va a considerar, se basará en el conocimiento que se tenga del fenómeno en el cual intervienen las variables que se están considerando.

Así por ejemplo el dominio de los valores que puede asumir la variable *X*, puede constituir una restricción en la definición de la familia de modelos, por tal razón es el especialista del área del estudio del fenómeno, quien dirá en primera instancia que familia considerar.

Por ejemplo, si se sabe por el comportamiento del fenómeno, que el crecimiento de *Y* por cada unidad que *X* crece, es constante, es decir:

$$\frac{dy}{dx} = k$$

Entonces la familia a considerar será $y = kx + c$ ó sea la familia de los modelos rectilíneos.

De esta manera será el agrónomo, el salubrista, el médico, el biólogo, etc. la persona que en primera instancia recomendará la familia de modelos a considerar, según sea el área de estudio, o proporcionará las pistas necesarias para proponer familias de modelos que sean razonables.

Si no se tuviera información sobre el fenómeno y se está en una etapa exploratoria, la forma del diagrama de dispersión puede sugerir el tipo de familia a considerar.

De esta manera y tomando el ejemplo de la familia de modelos rectilíneos, la preocupación sería entonces, encontrar entre las rectas la que "mejor" se ajuste a la nube de puntos.

El criterio que se usará para definir lo que se entiende por "el mejor modelo de la familia" es el **criterio de los mínimos cuadrados** y al modelo que satisfaga ese criterio se lo llamará **línea de regresión mínimo cuadrática.**

4.3.1 Criterio de los mínimos cuadrados

Se ilustra el criterio preliminarmente con un ejemplo sencillo, en el que se pretende ajustar una línea recta.

Ejemplo 4.1

El esfuerzo cortante del suelo en un cierto estrato arcilloso, parece estar relacionado con la profundidad.

En la región de “Igor” se toman 10 muestras de suelo a diferentes profundidades y se mide a cada una el esfuerzo cortante, en miles de libras por pie cuadrado [Klb/pie^2]. Se desea construir un modelo que permita hacer estimaciones del esfuerzo promedio del suelo que se encuentra a una profundidad de x pies.

CUADRO DE DATOS

OBSERVACION (i)	1	2	3	4	5	6	7	8	9	10

Profundidad *x* (pies)	6	8	14	14	18	20	20	24	28	30
Esfuerzo cortante *y* (Klb/pie^2)	0.28	0.58	0.50	0.83	0.71	1.01	1.29	1.50	1.29	1.58

Se sabe que no existe una asociación funcional perfecta entre profundidad y esfuerzo, es decir que puede suceder que diferente muestras que están a la misma profundidad, pueden tener distintas fuerzas cortantes, de hecho si miramos los datos esto se revela en las dos muestras que se tomaron a 14 pies y también en las que se tomaron a 20 pies de profundidad.

Sin embargo, las distribuciones de frecuencia del esfuerzo *y*, puede ser bien específica para el suelo que se encuentra a la misma profundidad *x*. En especial es de mucho interés encontrar un modelo que permita estimar la media *M(y/x)* para dicha distribución condicional de frecuencia.

M(y/x) es una función de *x*. Para hacerse una idea de la naturaleza de dicha función, de su forma, puede ser de mucha utilidad graficar en los puntos (*x*,*y*) en un plano cartesiano, dando origen al llamado " diagrama de dispersión ", como se muestra en el siguiente grafico:

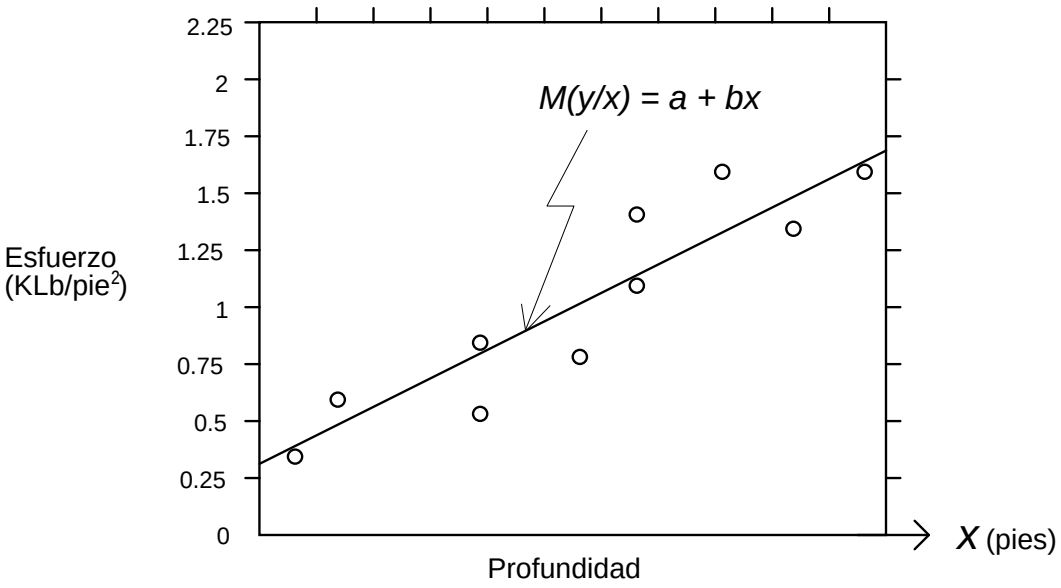


Fig. 4.2 a: Diagrama de dispersión del esfuerzo cortante y la profundidad.

A partir del diagrama de dispersión se puede se puede observar una cierta tendencia rectilínea de la nube de puntos, lo cual hace razonable pensar que el promedio *M(y/x)* tenga la forma de una línea recta, como se insinúa en el grafico:

$$M(y/x) = a + bx$$

Recordemos, que un buen indicador del grado de asociación de dos variables en la dirección de una línea recta es el coeficiente correlación lineal:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{S_x \cdot S_y}$$

para la situación del ejemplo se tiene que :

$\bar{x}$ = 18.2 pies $\bar{y}$ = 0.957 Klb/pie^2

S_x = 7.50733 pies S_y = 0.44385 Klb/pie^2

así que:

$$r_{xy} = 0.914$$

Es un valor alto, que significa que es muy razonable la propuesta de un modelo rectilíneo para *M(y/x).*

Queda ahora la tarea de hallar cual recta es. Es decir que cuales deben ser los valores de "a" y "b" que definen " la mejor " recta.

El criterio generalmente adaptado (no es el único criterio), para definir lo que significa " la mejor ", es el denominado criterio de los mínimos cuadrados (aunque debería decirse de los cuadrados mínimos).

En realidad, puede pensarse que para una observación (*x,y*) puede modelarse como:

$$y = M(y/x) + e$$

es decir que el valor del esfuerzo cortante para una observación particular tomada a una profundidad *x*, puede visualizarse como la media de su distribución condicional *M(y/x)* más lo que le haga falta, que hemos llamado *e* y se conoce como error.

De esta manera *e* es el error que se cometería si se quisiera predecir a *y*, con base en la media condicional *M (y/x)*, es decir:

$$e = y - M(y/x)$$

note que el error *e* puede ser de signo positivo o negativo.

El criterio de los mínimos cuadrados para encontrar " el mejor " modelo; consiste en calcular para cada posible modelo (en este caso rectas), los errores para los puntos

observados y en todas las posibilidades, seleccionar aquella que produce la menor suma de los errores al cuadrado.

Para el ejemplo, considerando el modelo $M(y/x) = a + bx$, definamos los errores para cada uno de los 10 puntos (x_i , y_i) que se observaron.

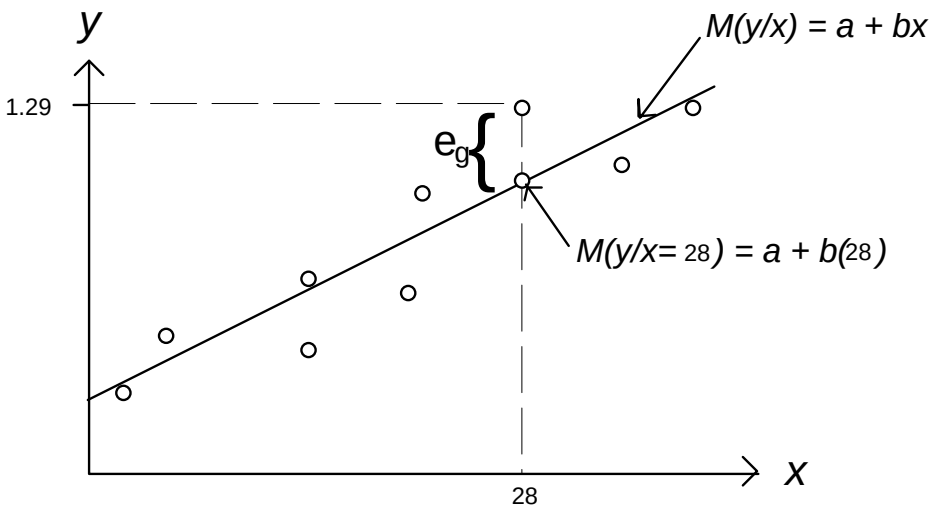


Fig. 4.2 b: Representación del error para una presentación preliminar.

Asi como muestra el gráfico: para el punto (28 , 1.29), el error asociado es

$$e_g = y_g - M(y/x_g)$$

$$= 1.29 - [a + b * 28]$$

note que si consideramos un modelo particular, “a” y “b” serian números conocidos y el error *e*, tendría por lo tanto un valor concreto.

Si hacemos este planteamiento para cada uno de los datos, se obtiene:

$e_1 = 0.28 - [a + b(6)]$

$e_2 = 0.58 - [a + b(8)]$

$e_3 = 0.50 - [a + b(14)]$

$e_4 = 0.83 - [a + b(14)]$

$e_5 = 0.71 - [a + b(18)]$

$e_6 = 1.01 - [a + b(20)]$

$e_7 = 1.29 - [a + b(20)]$

$e_8 = 1.50 - [a + b(24)]$

$e_9 = 1.29 - [a + b(28)]$

$e_{10} = 1.58 - [a + b(30)]$

El modelo queda perfectamente definido cuando se encuentren los numeros “a” y “b”. De todos los posibles, nos quedamos con aquellos que produzcan la menor suma:

$$e_1^2 + e_2^2 + e_3^2 + \ldots + e_{10}^2$$

note que dicha suma solo depende de los parámetros a y b del modelo, es decir que:

$$\sum_{i=1}^{10} e_i^2 = f(a,b) \qquad \leftarrow \quad \text{función de a y b.}$$

Aquí la situación se convierte en un problema de matemáticas: “hallar el mínimo cuadrado de *f (a , b)“* (para lo cual deben hallarse las derivadas y todo lo demás, que se tratará más adelante).

Ahora nos conformaremos con saber que al resolver el problema de minimizar nuestra función, resulto el siguiente sistema de ecuaciones lineales:

$$na + \left(\sum_{i=1}^{n} x_i\right) \cdot b = \sum_{i=1}^{n} y_i$$

$$\left(\sum x_i\right) \cdot a + \left(\sum_{i=1}^{n} x_i^2\right) \cdot b + \sum_{i=1}^{n} x_i y_i$$

Estas se conocen como **ecuaciones normales**.

Por ahora no se preocupe mucho por saber de donde salieron las ecuaciones. Expresemos el sistema de acuerdo a los datos concretos obtenidos en el problema.

De esta manera, las ecuaciones normales adoptan la forma:

$$10\,a + 182\,b = 9.57$$
$$182\,a + 3876\,b = 203.23$$

CUADRO 4.2

Observacion i	profundidad x_i	esfuerzo cortante y_i	$x_i\,y_i$	x^2_i	y^2_i	$M(y/x_i)$	e_i	e^2_i
1	6	0.28	1.68	36	0.078	0.325	-0.045	0.0020
2	8	0.58	4.64	64	0.336	0.429	0.151	0.0228
3	14	0.50	7.00	196	0.250	0.739	-0.239	0.0571
4	14	0.83	11.63	196	0.689	0.739	0.091	0.0083
5	18	0.71	12.78	324	0.504	0.946	-0.236	0.0557
6	20	1.01	20.20	400	1.020	1.049	-0.039	0.0015
7	20	1.29	25.80	400	1.662	1.049	0.241	0.0580
8	24	1.50	36.00	576	2.250	1.257	0.243	0.0590
9	28	1.29	36.10	784	1.662	1.463	-0.173	0.0299
10	30	1.58	47.40	900	2.495	1.566	0.014	0.0002
Σ	182	9.57	203.23	3876	10.946	9.57	0	0.2945
	↑	↑	↑	↑	↑	↑		↑
	$\Sigma\,x_i$	$\Sigma\,y_i$	$\Sigma\,x_i\,y_i$	$\Sigma\,x_i^2$	$\Sigma\,y_i^2$	$\Sigma\,e_i$		$\Sigma\,e_i^2$

Un sistema de dos ecuaciones lineales, con dos incógnitas, que al resolverlo resulta:

$$a = 0.015 \qquad b = 0.0517$$

Es decir que el modelo de regresión rectilíneo, obtenido con el criterio de los mínimos cuadrados es:

$$M(y/x) = 0.015 + 0.0517\,x$$

4.3.1.1 Como usar el modelo de regresión obtenido?

Que resultado arroja el modelo de regresión para $x = 10$ pies y que significa?

$$M(y/x = 10) = 0.015 + 0.0517\,(10)$$
$$= 0.532 \text{ Klb/pie}^2$$

Lo cual significa que para el suelo que se encuentra a una profundidad de 10 pies, se espera aproximadamente un esfuerzo cortante promedio de 0.532 Klb/pie^2.
El modelo permite hacer predicciones sobre el esfuerzo cortante promedio para la profundidad que se pida (dentro del rango de los valores observados para *x*, en este caso entre 6 y 30 pies).

ALGUNAS OBSERVACIONES IMPORTANTES

1. Note que en el modelo:

$$M(y/x) = a + b\,x$$

$$\frac{\partial M(y/x)}{\partial x} = b \qquad \text{ó lo que es lo mismo:}$$

$$M(y/x_o + 1) - M(y/x_o) = b$$

Lo cual significa que la pendiente del modelo rectilíneo, puede interpretarse, como la diferencia del esfuerzo cortante promedio de suelos con un pie de diferencia en profundidad.

En otras palabras, para el caso del ejemplo, se diría que el esfuerzo cortante promedio del suelo aumenta en 0.0517 Klb/pie^2 por cada pie que aumenta la profundidad.

2. Nótese que:

$M(y/x = 0) = a$, lo cual podría interpretarse, en el contexto del ejemplo, como que en la superficie (a cero profundidad) el suelo tiene una resistencia promedio de 0.015 Klb/pie^2 . Sin embargo se debe tener mucho cuidado, pues para que una interpretación como esta sea válida, es necesario que existan observaciones muy cerca del valor $x = 0$. Así pues en este ejemplo dicha interpretación no es correcta y en cambio podría visualizarse el intercepto "a" como una constante de ajuste del modelo.

3. Para un modelo rectilíneo $M(y/x) = a + bx$, la solución de mínimos cuadrados que resulta de despejar a y b de las ecuaciones normales, conduce a:

$$b = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum\left(x_i - \bar{x}\right)^2} = \frac{\mathrm{Cov}(x, y)}{S_x^2}$$

$$= \frac{S_{xy}}{S_x^2} = r\frac{S_y}{S_x}$$

$$\frac{\sum x_i\, y_i \; - \; n\bar{x}\bar{y}}{\sum x_i^2 \; - \; n(\bar{x})^2}$$ donde *r* es el coeficiente de correlación lineal

$$a \;=\; \bar{y} \;-\; b\,\bar{x}$$

Nótese que lo desarrollado en el ejemplo sólo es válido para la familia de modelos rectilíneos $M(y/x) = a + bx$, sin embargo, las ideas que se usaron para obtener los resultados siguen siendo válidos para cualquier otra familia de modelos, adaptando los criterios a las especificidades pertinentes.

En el ejemplo anterior, de antemano, se pudo obtener una idea de la calidad del modelo, usando como indicador de la expresión del coeficiente de correlación lineal, es importante resaltar que este indicador funcionaría solamente para la familia de modelos rectilíneos, $M(y/x) = a + bx$, para familias de modelos naturaleza distinta, se deberá desarrollar nuevos indicadores de la bondad de ajuste del modelo estimado.

A continuación se desarrolla en forma general el proceso de estimación de mínimos cuadrados, se explican sus alcances y limitaciones. Posteriormente se construye un indicador de bondad de ajuste de un modelo, aplicándole a un amplio espectro de modelos.

Con el propósito de simplificar la escritura, en algunas ocasiones se usará:

$$M\,(y/x) \;=\; y^* \;=\; f(x, ß)$$

Donde ß puede representar un conjunto de parámetros $ß_0, ß_1, ß_2, ..., ß_k$

4.3.1.2 Generalización de la estimación de parámetros de una familia de modelos usando el criterio de mínimos cuadrados

Se supone que se desea ajustar un modelo de la familia de la forma $Y^* = f(x,ß)$, donde ß representa un vector de parámetros $(ß_0, ß_1, ß_2, ..., ß_k)$; esto indica que cada juego de parámetros define de manera perfecta un modelo específico.

Se dispone de una muestra de *n* elementos a cada uno de los cuales se ha observado las característica *X* e *Y*, dando origen a los puntos: $(x_1, y_1), (x_2, y_2), ...,(x_n, y_n)$.

Si se usara el modelo $Y^* = f(x,ß)$, para predecir *Y*, en los elementos de la muestra, se tendría:

$$\begin{aligned} y^*_1 &= f(x_1,ß) \\ y^*_2 &= f(x_2,ß) \\ &\vdots \\ y^*_n &= f(x_n,ß) \end{aligned}$$

En general, estas predicciones no coinciden necesariamente con los valores observados de Y en la muestra $y_1, y_2,...,y_n$; esto implica que existen unos errores de predicción que para los distintos elementos de la muestra pueden escribirse como:

$$\begin{aligned} e_1 &= y_1 - y^*_1 \\ e_2 &= y_2 - y^*_2 \\ &\vdots \\ e_n &= y_n - y^*_n \end{aligned}$$

La magnitud de estos errores depende del modelo que se escoja, es decir, depende del juego de parámetros $ß = (ß_0, ß_1, ß_2, ...,ß_k)$ que se seleccione, como puede apreciarse en el gráfico de la figura 4.3.

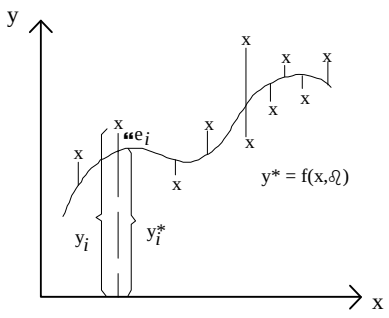


Fig. 4.3

Los trazos verticales que aparecen en la figura, corresponden a la magnitud de los distintos errores de predicción. Con el criterio de los mínimos cuadrados, se define el mejor modelo, entre los de una familia dada, como aquel que produzca la menor suma de los cuadrados de los errores de predicción.

El criterio de los mínimos cuadrados, como método para encontrar el mejor modelo de la familia $Y^* = f(x,ß)$, se puede expresar de la manera siguiente:

Encontrar $(ß_0, ß_1, ß_2, ..., ß_k)$ de tal manera que sea $e_1^2 + e_2^2 + ... + e_n^2$ la menor posible.

Si se parte del hecho, de que los datos $(x_1, y_1), (x_2, y_2), ...,(x_n, y_n)$ son conocidos entonces, la suma de los e_i^2 es una función de los $ß$.

$$\begin{aligned} e^2_1 &= [y1 - f(x_1 ß_0, ß_1, ..., ß_k)]^2 \\ e^2_2 &= [y_2 - f(x_2 ß_0, ß_1, ..., ß_k)]^2 \\ &\vdots \\ e^2_n &= [y_n - f(x_n ß_0, ß_1, ..., ß_k)]^2 \end{aligned}$$

$$\sum_{i=1}^{n} e_1^2 = \sum_{i=1}^{n} \left[y_i - f\left(x_i, \beta_0, \beta_1, \ldots, \beta_k\right)\right]^2 = G\left(\beta_0, \beta_1, \ldots, \beta_k\right)$$

De esta manera el método de los mínimos cuadrados consiste en aplicar la técnica de optimización adecuada para encontrar los *(ß$_0$, ß$_1$, ß$_2$, ..., ß$_k$)*, que hacen mínima la función $G(\beta_0, \beta_1, \beta_2, ..., \beta_k)$

Colocando a *f(x,ß)*, algunas condiciones, no muy restrictivas, puede resolverse el problema de:

hallar $ß_0, ß_1, ß_2, ..., ß_k$, que,

$$\min G(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} [y_i - f(x_i, ß_0, ß_1, ..., ß_k)]^2$$

Resolviendo el sistema:

$$\frac{\partial G(\beta)}{\partial \beta_0} = 0$$
$$\frac{\partial G(\beta)}{\partial \beta_1} = 0$$
.
.
.
$$\frac{\partial G(\beta)}{\partial \beta_k} = 0$$

Sistema de *(k+1)* ecuaciones con *(k+1)* incógnitas.

Si se tiene en cuenta que:

$$\frac{\partial(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} 2\left[y_i - f(x_i, \beta_0, \beta_1, ..., \beta_k)\right] . \frac{\partial f(x_i \beta)}{\partial \beta_j} \ , \ j = 0, 1, ... ,k$$

Entonces el sistema de ecuaciones puede escribirse como:

$$\frac{\partial G}{\partial \beta_0} = 2\sum_{i=1}^{n}\left[y_i - f\left(x_i,\ \beta_0,\ \beta_1, \ldots,\ \beta_k\right)\right].\frac{\partial f\left(x_i,\beta\right)}{\partial \beta_0} = 0$$

$$\frac{\partial G}{\partial \beta_1} = 2\sum_{i=1}^{n}\left[y_i - f\left(x_i,\ \beta_0,\ \beta_1, \ldots,\ \beta_k\right)\right].\frac{\partial f\left(x_i,\beta\right)}{\partial \beta_1} = 0$$

.

.

$$\frac{\partial G}{\partial \beta_k} = 2\sum_{i=1}^{n}\left[y_i - f\left(x_i,\ \beta_0,\ \beta_1, \ldots,\ \beta_k\right)\right].\frac{\partial f\left(x_i,\beta\right)}{\partial \beta_k} = 0$$

Este sistema de ecuaciones es conocido como ecuaciones normales, puede expresarse en forma más simplificada en términos del error de predicción:

$$e_i = y_i - f(x_i , ß_0 , ß_1 , ß_2 , \ldots , ß_k)$$

De esta manera, las ecuaciones normales son equivalentes a:

$$(0) \quad \sum_{i=1}^{n} e_i \frac{\partial f\left(x_i,\beta\right)}{\partial \beta_0} = 0$$

$$(1) \quad \sum_{i=1}^{n} e_i \frac{\partial f\left(x_i,\beta\right)}{\partial \beta_1} = 0$$

. ECUACIONES NORMALES (E 4.1)

.

$$(k) \quad \sum_{i=1}^{n} e_i \frac{\partial f\left(x_i,\beta\right)}{\partial \beta_k} = 0$$

No obstante el problema consiste en dar solución a un sistema de *(k+1)* ecuaciones con *(k+1)* incógnitas, esto no siempre es sencillo. En general si la función *f(x,ß)* es tal que el sistemas de ecuaciones no resulta lineal, entonces la situación se torna compleja. Cuando el sistema es lineal existen técnicas muy conocidas para su solución.

A continuación se analizan algunos casos de uso frecuente, sobre la forma de la familia de modelos $f(x,ß)$.

4.3.1.3 Caso en el cual la familia de modelos a considerar es lineal en los parámetros.

Es decir cuando $f(x ,ß_0, ß_1, ß_2, \dots , ß_k)$ es una **función lineal en los parámetros.**

Recuérdese que en las ecuaciones normales, se está considerando como variables a $ß_0, ß_1, ß_2, \dots , ß_k$ puesto que $(x_1, y_1), (x_2, y_2), \dots,(x_n, y_n)$ son datos conocidos, entonces la linealidad hace referencia a $ß_0, ß_1, ß_2, \dots, ß_k$. Así pues que en forma general una función lineal en los parámetros puede expresarse como:

$$f(x, ß_0, ß_1, \dots, ß_k) = ß_0 + ß_1 f_1(x) + \dots + ß_k f_k(x)$$

donde $f_1(x), f_2(x), \dots, f_k(x)$ son funciones que sólo dependen de x y no de los $ß$. Obsérvese que las $f_j(x)$ no tienen que ser necesariamente funciones lineales en x, pueden ser cualquier función; la única restricción es que no involucre los parámetros $ßj$ en su expresión, de esta manera, por ejemplo, la función:

$$f(x, ß_0, ß_1, ß_2) = ß_0 + ß_1 x^2 + ß_2 \ln x$$

es una función lineal en $ß_0, ß_1, ß_2$ en este ejemplo:

$f_1(x) = x^2$; $f_2(x) = \ln x$ que no constituyen funciones lineales en X.

Véase que ocurre entonces, con las ecuaciones normales, cuando $f(x,ß)$ es lineal en los parámetros, es decir cuando es de la forma:

$$f(x, ß_0, ß_1,\dots,ß_k) = ß_0 + ß_1 f_1(x) + \dots + ß_k f_k(x)$$

Obsérvese que en esta situación:

$$\frac{\partial f}{\partial \beta_0} = 1 \; ; \quad \frac{\partial f}{\partial \beta_1} = f_1(x), \dots, \frac{\partial f}{\partial \beta_k} = f_k(x)$$

Así, las ecuaciones normales (E 4.1) se convierten en:

$$
\begin{aligned}
&(0) \quad \sum_{i=1}^{n} e_i = 0 \\
&(1) \quad \sum_{i=1}^{n} e_i \cdot f_1(x_i) = 0 \\
&\quad . \\
&\quad . \\
&(k) \quad \sum_{i=1}^{n} e_i \cdot f_k(x_i) = 0
\end{aligned}
\qquad \text{(E 4.2)}
$$

Estas constituyen un sistema de *(k+1)* ecuaciones lineales con *(k+1)* incógnitas, el cual tiene solución muy definida por varios métodos, lo cual constituye una gran ventaja.

Se ilustra a continuación el proceso de estimación de los *ß's* que corresponden al mejor modelo de una familia dada de modelos lineales en los parámetros.

Ejemplo 4.1

Existe interés en determinar un modelo que permita "predecir" la resistencia de cierto tipo de concreto a los 28 días de curado, con base en la resistencia medida a los 10 días.

Con este propósito, se diseño un experimento que permitió para una muestra de 30 ensayos hacer las mediciones de resistencia de los 10 días (*X*) y los 28 días (*Y*), arrojando los siguientes resultados[3] expresados en libras/pulg^2. (psi):

Resistencia a los 10 días de "curado" *X*(psi)	Resistencia a los 28 días de "curado" *Y*(psi)
1800	2800
2135	2750
1450	2640
2140	2530
1870	2740
1945	2300
1720	2270

[3] Los resultados y las funciones propuestas en el ejemplo 4.1 no son reales sino hipotéticas.

2230	3040
Resistencia a los 10 días de "curado" *X*(psi)	Resistencia a los 28 días de "curado" *Y*(psi)
1540	3120
2100	2850
2400	3235
2650	3000
1765	2720
1280	2005
1350	1900
1980	2700
2000	3010
2380	3140
2070	2870
1990	2740
1775	2180
1748	2320
2135	2980
1534	2650
2320	3000
2188	3102
1831	2930
1302	2740
2005	2955
1434	2328

Estudios anteriores permiten pensar que la familia de modelos que pueden explicar estadísticamente el fenómeno es de la forma:

$$f(x) = ß_0 + ß_1 x + ß_2 x^2$$

Con base en el método de los mínimos cuadrados, plantee las ecuaciones normales y haga las estimaciones para $ß_0$, $ß_1$, $ß_2$, que corresponden al mejor modelo de la familia en estudio.

Como puede apreciarse la familia de modelos propuesta es lineal en los parámetros; de acuerdo con la expresión general de este tipo de modelos:

$$f(x, ß_0, ß_1, ß_k) = ß_0 + ß_1 f_1(x) + ß_2 f_2(x) + ... + ß_k f_k(x)$$

Significa que para la familia de modelos a estudiar

$$f_1(x) = x \quad ; \quad f_2(x) = x^2$$

Las ecuaciones normales de acuerdo con la expresión (E 4.2), serán:

$$(0) \quad \sum_{i=1}^{n} e_i = 0$$

$$(1) \quad \sum_{i=1}^{n} e_i \cdot f_1(x_i) = 0$$

$$(2) \quad \sum_{i=1}^{n} e_i \cdot f_2(x_i) = 0$$

Si se reemplaza $e_i = y_i - y_i^*$, donde:

$y_i^* = f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$, es decir:
$e_i = y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2$

Las ecuaciones normales pueden escribirse como:

$$(0) \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right) = 0$$

$$(1) \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right) x_i = 0$$

$$(2) \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right) x_i^2 = 0$$

Destruyendo los paréntesis, distribuyendo las sumatorias, y trasponiendo los términos que no están afectados por los *ß's*, se obtiene el sistema de ecuaciones lineales, expresado en su forma clásica.

$$(0) \quad \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum^{n} x_i + \beta_2 \sum^{n} x_i^2$$

$$(1) \quad \sum_{i=1}^{n} y_i x_i = \beta_0 \sum^{n} x_i + \beta_1 \sum^{n} x_i^2 + \beta_2 \sum^{n} x_i^3 \qquad \text{(E 4.3)}$$

$$(2) \quad \sum_{i=1}^{n} y_i x_i^2 = \beta_0 \sum^{n} x_i^2 + \beta_1 \sum^{n} x_i^3 + \beta_2 \sum^{n} x_i^4$$

Como se dispone de los datos (x_i,y_i), entonces las incógnitas en la ecuaciones (E 4.3), sólo son *ß*$_0$, *ß*$_1$, *ß*$_2$.

Los miembros de la izquierda constituyen constantes y las sumatorias de los términos de la derecha actúan como coeficientes de las incógnitas.

Haciendo los cómputos con los datos del ejemplo se obtiene:

$$n = 30\,; \quad \sum_{i=1}^{30} y_i = 81545\,; \quad \sum_{i=1}^{30} x_i = 57067$$

$$\sum_{i=1}^{30} x_i^2 = 1.12 \times 10^8\,; \quad \sum_{i=1}^{30} x_i^3 = 2.26 \times 10^{11}\,; \quad \sum_{i=1}^{30} x_i^4 = 4.69 \times 10^{14}$$

$$\sum_{i=1}^{30} y_i x_i^2 = 3.13 \times 10^{11}\,; \quad \sum_{i=1}^{30} y_i x_i = 1.57 \times 10^8$$

De acuerdo con esto, para el ejemplo, las ecuaciones normales quedan expresadas de la siguiente manera:

$$(0) \quad 81545 = 30ß_0 + 57067ß_1 + 1.12 \times 10^8 ß_2 \qquad \text{(E 4.4)}$$

$$(1) \quad 1.57 \times 10^8 = 57067ß_0 + 1.12 \times 10^8 ß_1 + 2.26 \times 10^{11} ß_2$$

$$(2) \quad 3.13 \times 10^{11} = 1.12 \times 10^8 ß_0 + 2.26 \times 10^{11} ß_1 + 4.69 \times 10^{14} ß_2$$

Al resolver el sistema (E 4.4), por cualquiera de los métodos existentes, se obtiene :

$$ß_0 = 4002 \qquad ß_1 = -2.00936 \qquad ß_2 = 0.00067994$$

lo cual significa que el modelo mínimo cuadrático es:

$$f(x) = 4002 - 2.00936x + 0.00067994x^2$$

así pues si $X = 1900$ psi, entonces:

$$y^* = f(1900) = 2638.8 \text{ psi}$$

Lo cual significa que se espera que para los ensayos en los cuales la resistencia a los 10 días es de 1900 psi, la resistencia promedia a los 28 días sea 2638.8 psi.

Debe recalcarse que la relación entre *X* e *Y* no es funcional, por tanto la predicción de *Y* con base *X*, se realiza a través de *M(Y/x)*, lo anterior puede escribirse:

$$y^* = M(Y/x = 1900) = 2638.8$$

Ejemplo 4.2

Con los mismos datos del ejemplo 4.1, se desea ajustar un modelo de la familia de los modelos rectilíneos, es decir, de la forma:

$$f(x) = ß_0 + ß_1 x$$

Como puede apreciarse también es una modelo lineal en los parámetros $ß_0$, $ß_1$,(aunque en este caso en especial, también es lineal en *x*).

En este caso $f_1(x) = x$, así que las ecuaciones normales de acuerdo con (E 4.2)

$$(0) \quad \sum_{i=1}^{n} e_i = 0$$

$$(1) \quad \sum_{i=1}^{n} e_i x = 0$$

Como $e_i = y_i - y_i^* = y_i - f(x_i)$

$$e_i = y_i - \beta_0 - \beta_1 x_i$$

haciendo el reemplazo de e_i, las ecuaciones normales quedan:

$$(0) \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$(1) \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

Destruyendo el paréntesis y distribuyendo las sumatorias, pueden expresarse de la forma clásica:

$$(0) \quad \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum x_i$$

$$(1) \quad \sum_{i=1}^{n} y_i x_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 \qquad \text{(E 4.5)}$$

De nuevo, al lado izquierdo quedan las constantes y las sumas del lado derecho representan los coeficientes de las incógnitas.

Evaluando dichas ecuaciones con los datos disponibles se obtiene:

$$\begin{aligned} (0)\ 81545 &= 30ß_0 + 57067ß_1 \\ (1)\ 1.57 \times 10^8 &= 57067ß_0 + 1.12 \times 10^8 ß_1 \end{aligned} \qquad \text{(E 4.6)}$$

Al resolver el sistema (E 4.6) se obtiene que:

$ß_0 = 1678.84 \quad ; \quad ß_1 = 0.54637$

Lo cual significa que el modelo rectilíneo mínimo cuadrático es:

$$f(x) = 1678.84 + 0.54637x$$

así, si $X = 1900$ psi , entonces:

$y^* = f(1900) = 2716.94$ psi

que debe interpretarse como la resistencia promedia a los 28 días para conjunto de ensayos para los cuales la resistencia a los 10 días fue de 1900 psi.

Ejemplo 4.3

Con los mismos datos del ejemplo 4.1, sobre resistencia de cierto tipo de concreto, se desea ajustar un modelo de la familia de la forma:

$$f(x) = \beta_0 + \beta_1 \ln x + \beta_2 \sqrt{x}$$

donde ln*x* es logaritmo natural de *x*.

Obsérvese que aunque la expresión de *f(x)* aparece ln*x* y también $\sqrt{x}$, el modelo es lineal en los parámetros *ß0, ß1, ß2*. De acuerdo con la expresión general de los modelos lineales:

$$f_1(x) = \ln x \quad ; \quad f_2(x) = \sqrt{x}$$

así pues, las ecuaciones

$$\begin{aligned} &(0) \quad \sum_{i=1}^{n} e_i = 0 \\ &(1) \quad \sum_{i=1}^{n} e_i \cdot f_1(x_i) = 0 \\ &(2) \quad \sum_{i=1}^{n} e_i \cdot f_2(x_i) = 0 \end{aligned} \qquad \text{(E 4.2)}$$

Teniendo en cuenta que:

$$e_i = y_i - \beta_0 - \beta_1 \ln x - \beta_2 \sqrt{x}$$

Destruyendo los paréntesis y distribuyendo las sumatorias, las ecuaciones normales se convierten en:

$$\begin{aligned} &(0) \quad \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} \ln x_i + \beta_2 \sum_{i=1}^{n} \sqrt{x_i} \\ &(1) \quad \sum_{i=1}^{n} y_i \ln x_i = \beta_0 \sum_{i=1}^{n} \ln x_i + \beta_1 \sum_{i=1}^{n} (\ln x_i)^2 + \beta_2 \sum_{i=1}^{n} \sqrt{x_i} \ln x_i \\ &(2) \quad \sum_{i=1}^{n} y_i \sqrt{x_i} = \beta_0 \sum_{i=1}^{n} \sqrt{x_i} + \beta_1 \sum_{i=1}^{n} \sqrt{x_i} \ln x_i + \beta_2 \sum_{i=1}^{n} \sqrt{x_i} \sqrt{x_i} \end{aligned}$$

Al calcular las diferentes sumatorias con base en el siguiente cuadro, que contiene respectivamente: el número de la observación, la resistencia a los 10 días (*X*), la raíz cuadrada de *X*, el logaritmo natural de *X*, y por último la resistencia a los 28 días (*Y*) que constituye la característica a predecir.

Obser-vación #	Resist. a los diez días	$\mathrm{Ln}(x)$	$\sqrt{x}$	esistencia a los 28 dias (y)	$[\mathrm{Ln}(x)]^2$	$(\sqrt{x})^2$	$\sqrt{x}\mathrm{Ln}x$	$y_i \mathrm{Ln}x_i$	$y_i\sqrt{x_i}$	$M(y/x_i)$	e_i	e_i^2
1	1800	7.5000	42.4300	2800	56.2500	1800.3050	318.2250	21000.0000	118804.0000	2680.4580	119.5422	14290.3500
2	2135	7.6700	46.2100	2750	58.8289	2135.3640	354.4307	21092.5000	127077.5000	2792.2940	-42.2937	1788.7600
3	1450	7.2800	38.0800	2640	52.9984	1450.0860	277.2224	19219.2000	100531.2000	2569.5060	70.4941	4969.4100
4	2140	7.6700	46.2600	2530	58.8289	2139.9880	354.8142	19405.1000	117037.8000	2795.4110	-265.4110	70443.0100
5	1870	7.5300	43.2400	2740	56.7009	1869.6980	325.5972	20632.2000	118477.6000	2709.1050	30.8946	954.4800
6	1945	7.5700	44.1000	2180	57.3049	1944.8100	333.8370	16502.6000	96138.0000	2733.5860	-553.5860	306457.5000
7	1720	7.4500	41.4700	2270	55.5025	1719.7610	308.9515	16911.5000	94136.9000	2657.0270	-387.0270	149789.8000
8	2230	7.7100	47.2200	3040	59.4441	2229.7280	364.0662	23438.4000	143548.8000	2826.1260	213.8737	45741.9600
9	1540	7.3400	39.2400	3120	53.8756	1539.7780	288.0216	22900.8000	122428.8000	2598.1220	521.8781	272356.8000
10	2070	7.6400	45.5000	2870	58.3696	2070.2500	347.6200	21926.8000	130585.0000	2769.8810	100.1193	10023.8700
11	1990	7.6000	44.6100	2740	57.7600	1990.0520	339.0360	20824.0000	122231.4000	2743.5300	-3.5297	12.4600
12	1775	7.4800	42.1300	2180	55.9504	1774.9370	315.1324	16306.4000	91843.4000	2676.3220	-496.3220	246336.0000
13	2100	7.6500	45.8300	2850	58.5225	2100.3890	350.5995	21802.5000	130615.5000	2783.1710	66.8293	4466.1500
14	2400	7.7800	48.9900	3235	60.5284	2400.0200	381.1422	25168.3000	158482.7000	2885.4890	349.5108	122157.8000
15	2650	7.8800	51.4800	3000	62.0944	2650.1900	405.6624	23640.0000	154440.0000	2967.8890	32.1115	1031.1500
16	1765	7.4800	42.0100	2720	55.9504	1764.8400	314.2348	20345.6000	114267.2000	2668.8410	51.1591	2617.2500
17	1280	7.1500	35.7800	2005	51.1225	1280.2080	255.8270	14335.8000	71738.9000	2520.8050	-515.8050	266055.2000
18	1350	7.2100	36.7400	1900	51.9841	1349.8280	264.8954	13699.0000	69806.0000	2536.9520	-636.9520	405707.9000
19	1980	7.5900	44.5000	2700	57.6081	1980.2500	337.7550	20493.0000	120150.0000	2743.9560	-43.9559	1932.1230
20	2000	7.6000	44.7200	3010	57.7600	1999.8780	339.8720	22876.0000	134607.2000	2750.3880	259.6122	67398.4900
21	2380	7.7700	48.7900	3140	60.3729	2380.4640	379.0983	24397.8000	153200.6000	2880.3040	259.6958	67441.9000
22	1748	7.4700	41.8100	2320	55.8009	1748.0760	312.3207	17330.4000	96999.2000	2663.6560	-343.6560	118099.4000
23	2135	7.6700	46.2100	2980	58.8289	2135.3640	354.4307	22856.6000	137705.8000	2792.2940	187.7063	35233.6500
24	1534	7.3400	39.1700	2650	53.8756	1534.2890	287.5078	19451.0000	103800.5000	2593.7580	56.2424	3163.2070
25	2320	7.7500	48.1700	3000	60.0625	2320.3490	373.3175	23250.0000	144510.0000	2856.2180	143.7819	20673.2400
26	2188	7.6900	46.7800	3102	59.1361	2188.3680	359.7382	23854.4000	145111.6000	2813.2630	288.7375	83369.3400
27	1831	7.5100	42.7900	2930	56.4001	1830.9840	321.3529	22004.3000	125374.7000	2695.6180	234.3819	54934.8600
28	1302	7.1700	36.0800	2740	51.4089	1301.7660	258.6936	19645.8000	98859.2000	2524.9410	215.0593	46250.5100
29	2005	7.6000	44.7800	2955	57.7600	2005.2480	340.3280	22458.0000	132324.9000	2754.1290	200.8714	40349.3200
30	1434	7.2700	37.8700	2328	52.8529	1434.1370	275.3149	16924.6000	88161.3600	2563.6980	-235.6980	55553.3200

Las ecuaciones normales que resultan son:

$$(0)\ 81545 = 30ß_0 + 226.02ß_1 + 1302.965ß_2$$

$$(1)\ 613867.4 = 226.02ß_0 + 1703.883ß_1 + 9820.885ß_2 \qquad \text{(E 4.7)}$$

$$(2)\ 3568212 = 1302.965ß_0 + 9820.885ß_1 + 57067ß_2$$

Cuya solución conduce a:

$$ß_0 = 5498.34 \quad ; \quad ß_1 = -728.432 \quad ; \quad ß_2 = 62.3464$$

lo cual significa que el modelo de regresión mínimo cuadrática es:

$$M(Y/x) = 5498.4 - 728.432\ \mathrm{Ln}x + 62.3464\sqrt{x}$$

Así pues si $X = 1900$ psi, entonces:

$$y^* = f(1900) = 2716.58 \text{ psi}$$

que representa una aproximación a *M(Y/x = 1900)* y que se usa en la predicción de *Y*.

Como puede observarse, en los ejemplos realizados la solución de las ecuaciones normales se ha reducido a la solución de un conjunto de m ecuaciones lineales con *m* incógnitas; esto ha ocurrido porque la familia de los modelos estudiadas han sido modelos lineales en los parámetros; de no ser así en la solución de las ecuaciones normales surgen problemas adicionales que en ocasiones es necesario resolver con métodos numéricos, haciendo uso de medios iterativos con la ayuda de un computador.

Hasta ahora se conoce la técnica para ajustar un conjunto de datos (X_i,Y_i) el mejor modelo entre los de una familia dada de modelos lineales en los parámetros, usando el método de los mínimos cuadrados.

Como se expresó inicialmente, el propósito de la construcción de modelos de regresión es poder realizar "predicciones" **confiables**.
Hace falta entonces definir entonces un instrumento que sirva de indicador, sobre la bondad del modelo encontrado, con base en el grado de ajuste del mismo a los datos.

4.4 INDICADOR DE LA BONDAD DE UN MODELO DE REGRESION

Como puede intuirse del gráfico de la figura 4.3, el modelo se ajusta de forma perfecta cuando todos los e_i son cero, o en forma equivalente, todas la predicciones y^*_i , para los distintos x_i de la muestra, coinciden en forma perfecta con los diferentes valores de y_i, observados .

Se sabe que el modelo de regresión mínimo cuadrático encontrado, produce la menor suma de los cuadrados de los errores. No obstante se sabe que $\sum e_i^2$ es la mínima , no se puede juzgar si es "pequeña" o "grande". Mirando la situación desde otro punto de vista, es posible también que si se ensayara otra familia de modelos, el mejor modelo de ésta, produzca una suma $\sum e_i^2$ mínima, menor que la mínima de la primera familia, lo cual estaría indicando, de acuerdo con este criterio, que el segundo modelo es mejor que el primero, pero aún así no se sabe si es bueno o no en términos de la predicción, puesto que hasta ahora no se ha encontrado una cota para $\sum e_i^2$, que permita definir una escala.

Intuitivamente puede deducirse que una cota para la suma de los cuadrados de los errores, $\sum e_i^2$, está dada por $\sum\left(y_i - \bar{y}\right)^2$. A continuación se justifica esta exploración intuitiva.

$$\sum e_i^2 = \sum\left[y_i - f\left(x_i, \beta\right)\right]^2$$

Se espera que el peor de los casos, ocurra cuando la información que aporta la característica *X*, no ayude nada en la predicción de *Y*, lo cual significa que

$$y^* = M(y/x) = C \quad \text{constante}$$

En este caso, como de acuerdo con las ecuaciones normales, para modelos lineales en los parámetros debe cumplirse que

$\sum e_i = 0$ entonces $\Sigma (y_i - M(y/x_i)) = 0$

y si $M(y/x) = C$ ==> $\Sigma (y_i - C) = 0$

==> $C = \dfrac{\sum y_i}{n} = \bar{y}$, lo cual significa que si *M(y/x)* es una constante ella debe ser $\bar{y}$.

Así pues que en esta situación extrema, $\sum e_i^2 = \sum \left(y_i - \bar{y}\right)^2$, de donde se desprende en general, para cualquier familia de modelos se cumple que:

$$0 \leq \sum e_i^2 \leq \sum \left(y_i - \bar{y}\right)^2 \qquad \text{(E 4.8)}$$

Obsérvese que para un conjunto de datos, $\sum \left(y_i - \bar{y}\right)^2$ es un valor fijo que no depende de la familia de modelos que se desee estudiar, por tanto la expresión (E 4.8) constituye una escala que permite interpretar la magnitud de $\sum e_i^2$. De acuerdo con esto y teniendo en cuenta que cuando $\sum e_i^2 = 0$ el modelo se ajusta perfectamente a los datos observados y sabiendo que por (E 4.8):

$$0 \leq \frac{\sum e_i^2}{\sum \left(y_i - \bar{y}\right)^2} \leq 1$$

Puede definirse el **coeficiente de determinación**

$$R^2 = 1 - \frac{\sum e_i^2}{\sum \left(y_i - \bar{y}\right)^2}$$

De esta manera:

$$0 \leq R^2 \leq 1$$

Siendo $R^2 = 0$ cuando $\sum e_i^2 = \sum \left(y_i - \bar{y} \right)^2$ es decir, cuando x, no aporta información para predicción de y; por otro lado $R^2 = 1$ es decir, cuando $\sum e_i^2 = 0$ es decir, cuando el modelo mínimo cuadrático se ajusta en forma perfecta a los puntos (x_i, y_i) observados. En general el modelo será mejor, cuando más cerca de uno (1) esté el valor de R^2 correspondiente.

Con relación a los ejemplos anteriores, sus coeficientes de correlación R^2 son los siguientes:

Para la situación planteada en el ejemplo 4.2, donde se uso la variable de resistencia del concreto a los diez (10) días de curado (X), para predecir la resistencia a los 28 días de curado (Y) a través del modelo lineal:

$$M(y/x) = 1678.84 + 0.54637\, x$$

El coeficiente de determinación:

$$\sum_{i=1}^{30} \left(y_i - \bar{y} \right)^2 = 3521162 = \text{ variación total}$$

$$\sum e_i^2 = \sum_{i=1}^{n} \left[y_i - M\left(y / x \right) \right]^2 = 2079642 = \text{ variación residual}$$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum \left(y_i - \bar{y} \right)^2} = 1 - \frac{2079642}{3521162}$$

$$R^2 = 0.4094$$

lo cual significa que el modelo encontrado explica aproximadamente el 40.94% de la variación de Y en la muestra.

Aunque no existe una frontera para clasificar con base en R^2 los modelos en buenos y malos, puede decirse que este modelo no sería del todo confiable en la predicción de Y.

Por esta razón cuando un modelo de regresión simple (una sola variable predictiva x), el coeficiente de determinación no es muy alto, debe explorarse la situación para vincular mas variables al modelo con el propósito de explicar mayor porcentaje de la variación de y.

Para la situación planteada en el ejemplo 4.3, para predecir la resistencia a los 28 días de curado (Y) a través del modelo:

$$M(y/x) = 5498.4 - 728.432\,\mathrm{Ln}(x) + 62.346\sqrt{x}$$

El coeficiente de determinación:

$$\sum_{i=1}^{30}\left(y_i - \bar{y}\right)^2 = 3521162 = \text{ variación total}$$

$$\sum e_i^2 = \sum_{i=1}^{n}\left[y_i - M\left(y/x\right)\right]^2 = 2401138 = \text{ variación residual}$$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum\left(y_i - \bar{y}\right)^2} = 1 - \frac{2401138}{3521162}$$

$$R^2 = 0.318$$

Esto significa que el modelo encontrado explica aproximadamente el 31.8% de la variación de *Y* en la muestra. Puede decirse que este modelo no sería muy confiable en la predicción de *Y*.

Enseguida va a demostrarse que esas expresiones intuitivas tienen verdadero fundamento.

4.4.1 Expresión del análisis de varianza asociado a un modelo de regresión

Va a demostrarse que para una familia de modelos lineales en los parámetros, se cumple que para el modelo mínimo cuadrático, la expresión:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}\left[M\left(y/x_i\right) - \bar{y}\right]^2 \quad \text{(E 4.9)}$$

Donde:

$$M(y/x) = ß_0 + ß_1 f_1(x) + ß_2 f_2(x) + \ldots + ß_k f_k(X)$$

En la expresión (E 4.9), los términos:

$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ se conoce como la **variación total** y sólo depende de los datos, no

depende del modelo que se quiere ajustar, es decir que para un conjunto de datos dado, la variación total es una constante (el numerador de la varianza de *y*).

$\sum_{i=1}^{n} e_i^2$ Se conoce como la **variación residual** (ó variación no explicada), puesto que es la variación que permanece aún después de ajustar el modelo mínimo cuadrático, es decir la variación no explicada por el modelo. Evidentemente depende de la familia de modelos que se esté ajustando.

$\sum_{i=1}^{n}\left[M(y/x_i) - \bar{y}\right]^2$ Se conoce como la **variación explicada** por el modelo *M(y/x)*.

Variación total = variación explicada por *M(y/x)* + variación residual

Así, para que la suma sea constante, debe suceder que si la variación explicada aumenta, entonces la variación residual disminuya y viceversa .

Para probar la expresión (E 4.9), se parte del supuesto de que la familia de modelos que se estudia es lineal en los parámetros, es decir de la forma :

$$f(x) = \beta_0 + \beta_1 f_1(x) + \beta_2 f_2(x) + \ldots + \beta_k f_k(x)$$

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_i - M(y/x_i) + M(y/x_i) - \bar{y}\right)^2$$

$$= \sum_{i=1}^{n}\left[\left(y_i - M(y/x_i) + M(y/x_i) - \bar{y}\right)\right]^2$$

Recordando que $e_i = y_i - M(y/x_i)$ puede escribirse:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left[e_i + \left(M(y/x_i) - \bar{y}\right)\right]^2$$

elevando al cuadrado del binomio que está dentro de los corchetes se obtiene distribuyendo la sumatoria:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}\left[M(y/x_i) - \bar{y}\right]^2 + 2\sum_{i=1}^{n} e_i\left[M(y/x_i) - \bar{y}\right]^2$$

para obtener la expresión (E 4.9) que se desea probar, sólo restaría mostrar que el doble producto es cero, lo cual se logra recordando que las ecuaciones normales que dieron origen al modelo mínimo cuadrático *M(y/x)* son:

$$(0) \quad \sum_{i=1}^{n} e_i = 0$$

$$(1) \quad \sum_{i=1}^{n} e_i M_i(y / x_i) = 0$$

.

.

$$(k) \quad \sum_{i=1}^{n} e_i M_k(y / x_i) = 0$$

Por lo tanto:

$$2\sum e_i\left[M(y / x_i) - \bar{y}\right] = 2\left[\sum e_i M(y / x_i) - \bar{y}\sum e_i\right]$$

Pero

$$\begin{aligned}\sum_{i=1}^{n} e_i M(y / x_i) &= \sum e_i\left[\beta_0 + \beta_1 M_1(y / x_i) + \ldots + \beta_k M_k(y / x_i)\right] \\ &= \beta_0 \sum e_i + \beta_1 \sum e_i M_i(y / x_i) + \ldots + \beta_k \sum e_i M_k(y / x_i) \\ &= 0\end{aligned}$$

De esta manera se ha probado que si *M(y/x)* es el modelo mínimo cuadrático de una familia de modelos lineales en los parámetros entonces se cumple que:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}\left[M(y / x_i) - \bar{y}\right]^2$$

Expresión a partir de la cual puede construirse el indicador de bondad de ajuste que se mencionó anteriormente, conocido como **coeficiente de determinación.**

$$R^2 = \frac{\text{Variación explicada por } M(y/x)}{\text{Variación total}}$$

$$R^2 = \frac{\sum\left[M(y/x_i) - \bar{y}\right]^2}{\sum\left(y_i - \bar{y}\right)^2}$$

O en forma equivalente

$$R^2 = 1 - \frac{\text{Variación Residual}}{\text{variación total}}$$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum\left(y_i - \bar{y}\right)^2}$$

Obviamente $0 \leq R^2 \leq 1$

Ejemplo 4.4

Para la situación planteada en el ejemplo 4.1, donde se uso la variable de resistencia de concreto a los 10 días de curado (X), para predecir la resistencia a los 28 días de curado (Y) a través del modelo mínimo cuadrático:

$$M(y/x) = 4002 - 2.00936\,x + 0.00067994\,x^2$$

Usando el coeficiente de determinación conceptuar sobre la bondad del modelo hallado.

Como puede apreciarse de la expresión de R^2 es necesario calcular para cada x_i, la correspondiente estimación $M(y/x_i)$, por tal razón se construye el siguiente cuadro:

x_i(psi)	$y^*_i = M(y/x_i)$	y_i(psi)	e_i	e_i^2
1800	2588.158	2800	211.842	44877.200
2135	2811.336	2750	-61.336	3762.093
1450	2518.002	2640	121.998	14883.550

2140	2815.823	2530	-285.823	81694.690
1870	2622.179	2740	117.821	13881.790
1945	2666.025	2300	-486.025	236220.100
1720	2557.435	2270	-287.435	82619.050
2230	2902.401	3040	137.599	18933.530
1540	2520.131	3120	599.869	359842.500
2070	2756.099	2870	113.900	12973.280
1990	2696.004	2740	43.996	1935.649
1775	2577.622	2180	-397.622	158103.200
2100	2780.880	2850	69.121	4777.657
2400	3095.990	3235	139.010	19323.670
2650	3452.075	3000	-452.075	204371.500
1765	2573.646	2720	146.354	21419.590
1280	2544.033	2005	-539.033	290556.500
1350	2528.555	1900	-628.555	395080.900
1980	2689.104	2700	10.896	118.723
2000	2703.040	3010	306.960	94224.440
2380	3071.175	3140	68.825	4736.834
1748	2567.198	2320	-247.198	61106.910
2135	2811.336	2980	168.664	28447.58
1534	2519.646	2650	130.353	16992.000
2320	2999.994	3000	0.006	0.000038
2188	2860.627	3102	241.373	58260.930
1831	2602.402	2930	327.5978	107320.300
1302	2538.450	2740	201.550	40622.29
2005	2706.609	2955	248.391	61698.090
1434	2518.777	2328	-190.776	36395.660

De acuerdo con el cuadro anterior y con los datos obtenidos se obtiene que:

$$\sum_{i=1}^{30}\left(y_i - \bar{y}\right)^2 = 3521162 = \text{ variación total}$$

$$\sum e_i^2 = \sum_{i=1}^{n}\left[y_i - M\left(y/x\right)\right]^2 = 2372934 = \text{ variación residual}$$

Así:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum\left(y_i - \bar{y}\right)^2} = 1 - \frac{2372934}{3521162}$$

$$R^2 = 0.326$$

lo cual significa que el modelo encontrado explica aproximadamente el 32.6% de la variación de *Y* en la muestra.

Aunque no existe una frontera para clasificar con base en R^2 los modelos en buenos y malos, puede decirse que este modelo no sería muy confiable en la predicción de *Y*.

Aunque la aceptación de un modelo para la predicción, con base en el coeficiente de determinación, depende de los objetivos del modelo y la precisión requerida, puede decirse en forma muy general que modelos con $R^2 > 0.80$ pueden considerarse como relativamente buenos.

Es de anotar que en la complejidad de la naturaleza, se da con mucha frecuencia que la variabilidad de una característica *y*, es explicada por varias características.
Por esta razón cuando un modelo de regresión simple (una sola variable predictiva *x*), el coeficiente de determinación no es muy alto, muy probablemente debe explorarse la situación para vincular mas variables al modelo con el propósito de explicar mayor porcentaje de la variación de *y*. Esto da origen a los llamados modelos de regresión múltiple.

Por otro lado, no siempre es posible modelar los fenómenos con familias de modelos lineales en los parámetros, siendo forzoso usar familias de modelos no lineales, con las consiguientes dificultades que llevan inherentes.

4.4.2. Acerca de las familias de modelos no lineales en los parámetros.

Cuando se trató el método de los mínimos cuadrados, como una técnica para obtener el modelo de una familia que mejor se ajuste a un conjunto de puntos dados, se desarrolló en forma general para cualquier familia de modelos $f(x)$ y se plantearon en forma general las llamadas ecuaciones normales.

Se hizo notar que las ecuaciones normales tomaban la forma de un sistema de ecuaciones, de fácil solución cuando la familia de modelos a estudiar, es lineal en los parámetros. Se mencionó que cuando esto no ocurre la solución del sistema de ecuaciones normales es más complicado y que inclusive puede llegar a ser necesario el uso de métodos numéricos iterativos con ayuda del computador.
No obstante las dificultades que precedan el hallazgo del modelo mínimo cuadrático de una familia de modelos no lineales en los parámetros, existe un problema adicional: el juicio sobre su bondad, porque la expresión del análisis de la varianza asociado a la regresión se satisface cuando los modelos son lineales en los parámetros y como se vio, ésta expresión es la base para la definición del coeficiente de determinación. Resumiendo, este indicador no aplica en modelos no lineales en los parámetros.

No obstante, que esta situación restringe el campo de acción de los modelos no lineales, son de muy frecuente estudio algunos casos de modelos no lineales en los parámetros pero que son "linealizables" mediante alguna transformación, también es práctica generalizada que para esta clase de modelos se juzgue su bondad con base en el modelo linealizado, puesto que para el original el coeficiente de determinación no aplica, esta práctica debe usarse con reserva, **puesto que no es evidente la asociación entre la bondad del modelo linealizado y el original**. El proceso de linealización se ejecuta para facilitar la estimación de parámetros del modelo.

A continuación se presentan algunas familias de modelos linealizables y se hace explícita la transformación adecuada. El desarrollo del proceso de estimación de los parámetros del modelo, a partir del modelo linealizado, no se presenta, pues coincide con los desarrollados con el modelo lineales en los parámetros.

Modelos de la forma: $M(y/x) = \beta_0 x^{\beta_1}$

Puede aplicarse la transformación logarítmica; de esta manera:

$$Ln\ M(y/x) = Ln ß_0 + ß_1 lnx$$

si se hace:

$Ln\ M(y/x) = W$
$Lnx = T$
$Ln ß_0 = B_0$
$ß_1 = B_1$

asi si:

$$y_i = M(y/x) \cdot e_i$$
$$Ln\ y_i = Ln\,[M(y/x) + Ln\ e_i]$$
$$W_i = Ln\ ß_0 + ßLnx + e_i^*$$

se tiene: $W = B_0 + B_1 T$ que es un modelo lineal en B_0 y B_1

Modelos de la forma : $M(y/x) = ß_0\, ß_1^x \cdot e$

puede aplicarse:

$$ln\ M(y/x) = ln ß_0 + (ln ß_1)x + Ln\ e_i$$
$$W = ß_0 + ß\,x + e_i^*$$

Si se hace:

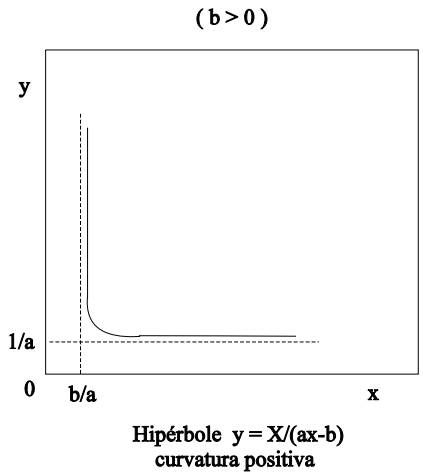
$LnM(y/x) = W$

$Ln\beta_0 = B_0$
$Ln\beta_1 = B_1$

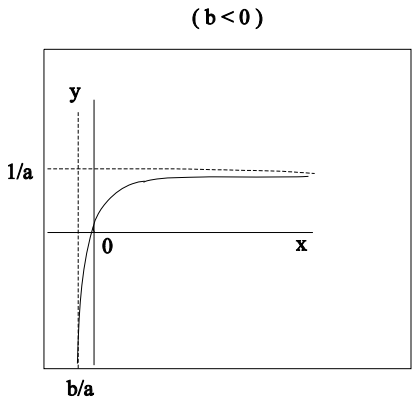
se tiene: $M(w/x) = B_0 + B_1 x$ que corresponde a un modelo lineal en B_0 y B_1.

4.4.2.1 OTROS MODELOS NO LINEALES EN LOS PARAMETROS

Cuando no se dispone de un modelo teórico que permita la estimación de los parámetros, es necesario identificar algunas posibilidades con base en los diagramas de dispersión.
A continuación representan algunas familias de curvas que pueden ser de utilidad al momento de la identificación. Las curvas que se presentan corresponden a modelos no lineales en los parametros pero que son linealizables por medio de una transformación.

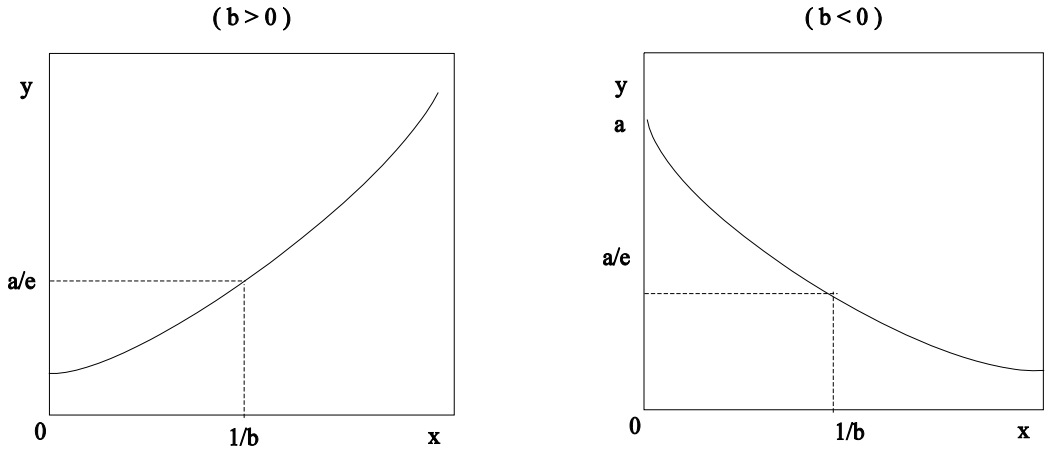


Hipérbole y = X/(ax-b)
curvatura positiva



Hipérbola y = X/(ax-b)
curvatura negativa

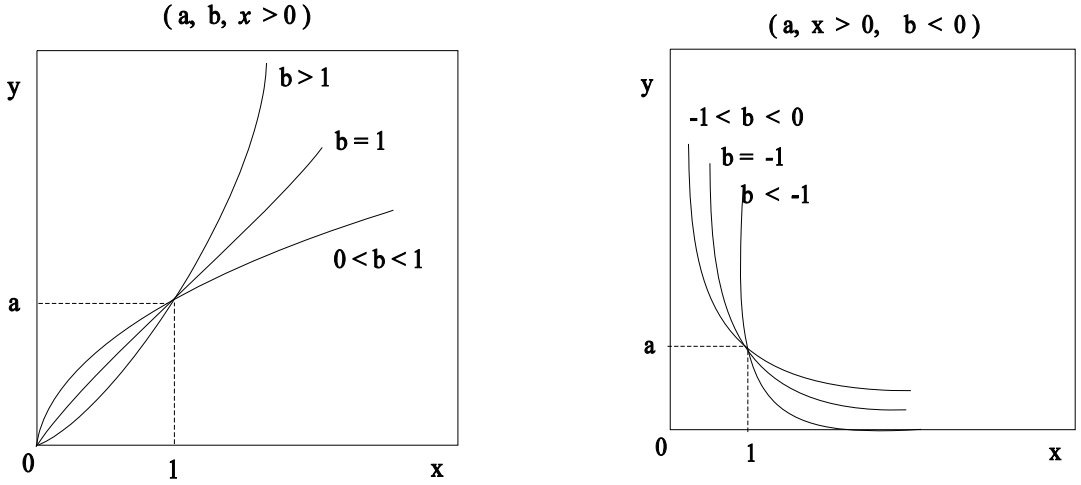
Forma lineal : $1/y = a - b/x$



Funciones exponenciales

$$Y = a\, e^{b\,x}$$

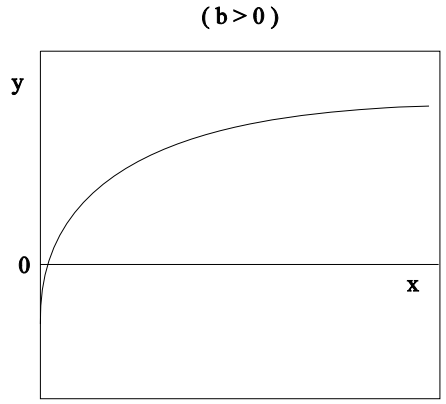
Forma Lineal : $LnY = Ln\, a + b\, X$

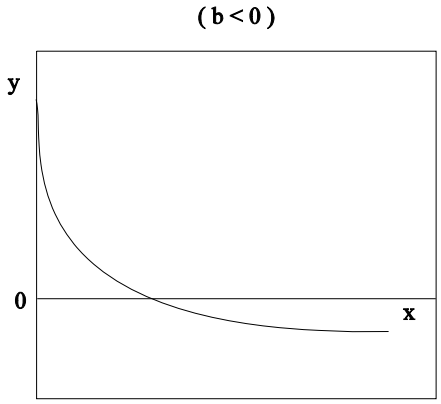


Funciones potenciales

$$Y = a\, x^{b}$$

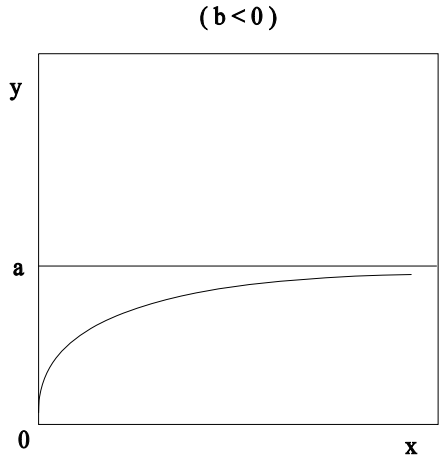
Forma Lineal : $LnY = Ln\, a + b\, Ln\, x$

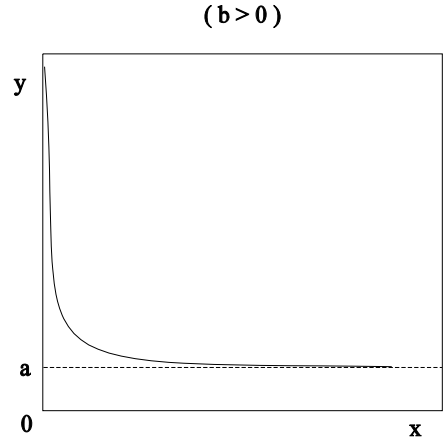




Funciones logarítmicas

En forma lineal : $y = a + b\, Ln\, x$

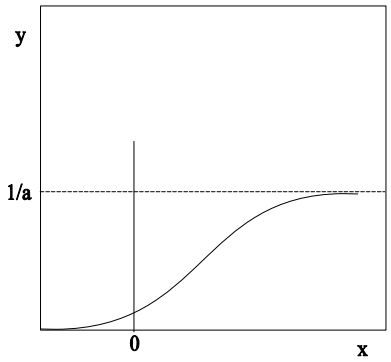




Funciones especiales

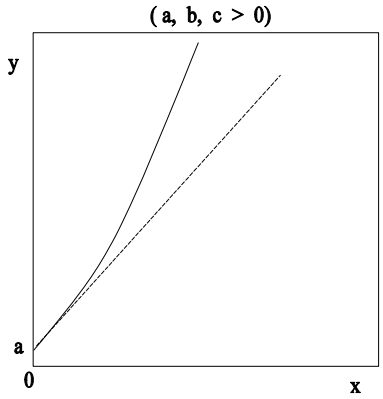
$$Y = a\, e^{b/x}$$

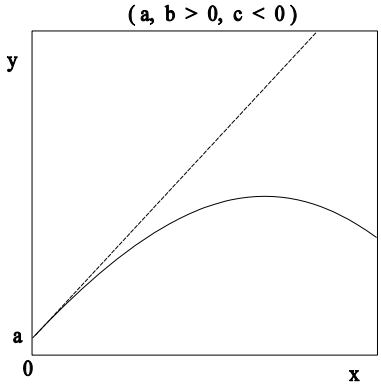
Forma Lineal : $LnY = Ln\, a + b/x$



$$Y = 1/(a + b\ e^{-x})$$

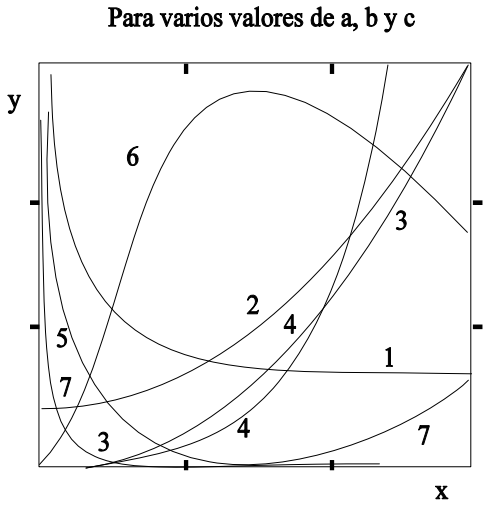
Forma Lineal : $1/Y = a + b\,e^{-x}$





Funciones polinomicas

Forma Lineal : $Y = a + b\,e^{-x}$



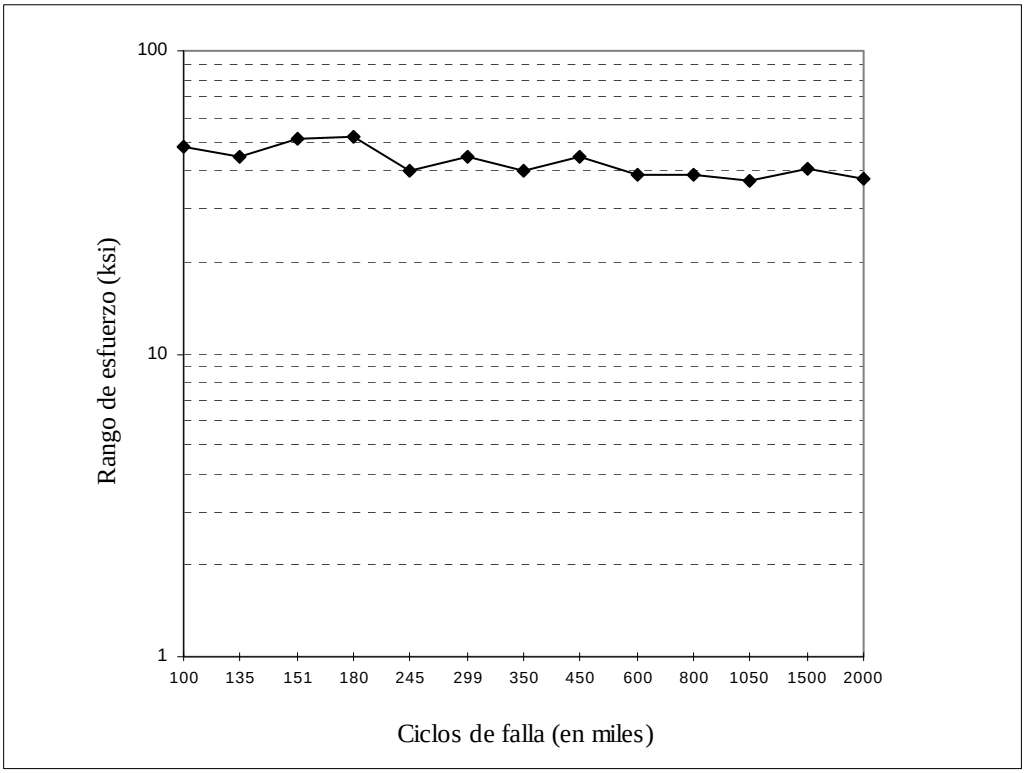
Funciones especiales de Hoerl

$y = a\ X^{b}\ e^{c\,x}$

Forma Lineal : $Ln\ y = Ln\ a + b\ Ln\ x + cx$

4.4.2.2 APLICACIONES DE UN ANALISIS DE REGRESION EN INGENIERIA

Los análisis de regresión son usados de modo muy práctico en todas las ramas de la ingeniería para obtener relaciones empíricas entre dos (o más) variables. Algunas veces la relación entre dos variables en ingeniería no puede deducirse con base en consideraciones teóricas; en estos casos la relación requerida entre las variables puede ser obtenida empíricamente con base en las observaciones experimentales. Por ejemplo para graficar el logaritmo de las observaciones de fatiga *N* de un material versus el logaritmo aplicado al rango de stress S, se observa una tendencia lineal asi como se muestra en la siguiente figura.



Esta tendencia se puede representar por

$$Log\ N = Log\ a - b\ Log\ S$$

La línea de regresión de *Log N* sobre *Log S* daría entonces las constantes a y b. Esta ecuación de regresión además sugiere una relación *S* - *N* de la forma

$$N S^{b} = a$$

En otras situaciones la forma matemática de requerimiento de vínculos quizás se deriva o postula de consideraciones fisicas; el análisis de regresión puede entonces ser usado para determinar los valores de los parámetros, o para evaluar la validez de la ecuación teórica.

4.5. SOBRE EL USO DE LOS MODELOS DE REGRESION

Es menester hacer algunas precisiones acerca del uso de las líneas de regresión.

- El modelos de regresión sólo puede usarse para hacer predicciones en el recorrido que la variable predictora tiene en los datos usados para obtener el modelo, es decir, **sólo se permite interpolar y no extrapolar**.

  En caso de que se use el modelo para extrapolar, a la predicción obtenida no puede asociarse ningún tipo de confianza estadística; en esta situación es el profesional del área específica que por su conocimiento del fenómeno en estudio, asume el riesgo de la extrapolación. En la figura que aparece a continuación se ilustra el riesgo de extrapolación.

  En el gráfico de la figura 4.4 la línea continua representa el modelo construido en el rango de datos y las líneas punteadas representan distintas alternativas para el curso de acción del fenómeno en la región donde no se tomó información, lo cual pone de manifiesto lo aventurado de la extrapolación.

- No debe olvidarse que los *ß's* que resultan al aplicar el criterio de los mínimos cuadrados, se ejecuta con base en una muestra, lo cual permite intuir que si se tomara otra muestra de la misma población los resultados podrían ser distintos, es decir existe una incertidumbre cuya magnitud puede estar asociada con el tamaño de la muestra, entre otras características. Existen herramientas en la inferencia estadística para cuantificar esta incertidumbre.

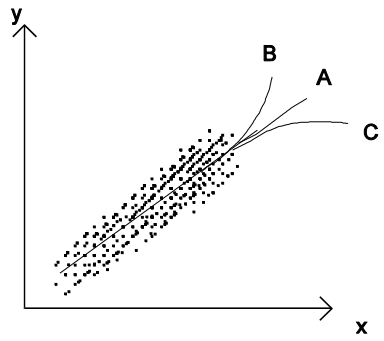


Fig. 4.4 Riesgo de extrapolación

  En la realidad, la complejidad de la mayoría de los fenómenos es tal que es difícil lograr explicar estadísticamente la variación de una característica, usando solamente otra.

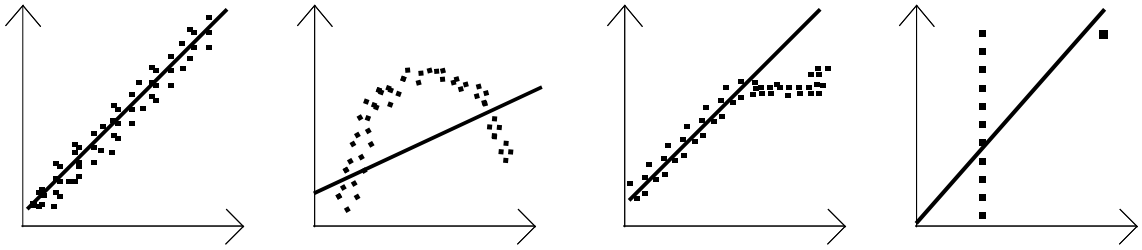
Con la misma base conceptual es posible generalizar los procedimientos desarrollados para la construcción de modelos que permitan involucrar varias variables en la explicación de cierta característica de interés.

4.5.1 Los supuestos del modelo de regresión

- Los desarrollos que se han presentado son todos de naturaleza exploratoria, sin embargo, como se discutió desde el principio, los resultados pueden variar de muestra en muestra. Surge aquí la pregunta, entonces ¿cómo creer en los resultados que provienen de una muestra, si para otra muestra los resultados no coinciden exactamente ?. La respuesta tiene varias aristas; la primera: la regularidad estadística, hace que a medida que la muestra se incrementa en su tamaño, la variación de muestra a muestra, sea cada vez menor, de tal manera que con una muestra suficientemente grande, tenemos gran confianza que los resultados puntuales obtenidos, no cambiarían mucho si se repitieran de nuevo el experimento o el estudio según sea el caso. La segunda arista, es que para cualquier tamaño de muestra, no necesariamente grande, es posible hacer afirmaciones probabilísticas acerca de los parámetros estimados y aún de las predicciones realizadas con el modelo, siempre y cuando se satisfagan ciertas condiciones o supuestos, que exige el modelo para realizar ese tipo de inferencias. Algunas de ellas son las siguientes:

- **Homogeneidad de Varianza**. La varianza de la distribución condicional de variable dependiente Y, debe ser constante, para cualquier valor de la variable independiente o predictora X. Cuando esto no se cumple, los estimadores de mínimos cuadrados ordinarios, no producen los mejores estimadores, razón por la cual deben realizarse algunas ponderaciones que corrijan este efecto. En el caso que ilustra en la figura, se nota que a medida que la variable X toma valores mas grandes, la variabilidad de la variable Y se hace mayor, es decir No se cumple la condición de “homogeneidad de varianza” y por el contrario se dice que hay “heterocedasticidad”.

- **Modelo adecuado**. Otra condición que se exige, es que el modelo propuesto sea el adecuado, lo cual significa que en realidad el modelo poblacional,
contenga las medias condicionales M(Y/x), para todos los valores de la variable predictora X. A continuación se muestran algunos casos en los que esta condición aparentemente obvia, no se satisface.



a) parece que el modelo es apropiado

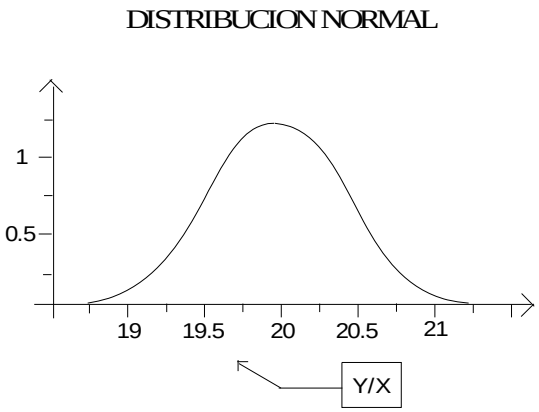
b) parece que el modelo rectilíneo es incorrecto, sugiere un modelo cuadrático

c) parece que el modelo rectilíneo es correcto para buena parte de los puntos;:

d)No hay información para juzgar el modelo, la pendiente está total y definida por el punto lejano

- **Independencia** de las n observaciones de la variable dependiente Y. Lo cual se puede garantizar, seleccionado la muestra de manera aleatoria (al azar).

- **La distribución** Condicional de Y, para cada realización de la Variable predictora X, debe ser aproximadamente "Normal", (campana de Gauss). como se muestra en la figura. note que corresponde a la distribución de Y para un valor particular de X=x.



Existen pruebas estadísticas para estar razonablemente seguro, que el modelo satisface aproximadamente los supuestos mencionados. Estas pruebas se escapan del objetivo de este libro, sin embargo en las referencias bibliográficas en la parte final de éste, se citan varios libros donde éstas se encuentran desarrolladas.

RELACIÓN ENTRE EL NÚMERO DE DATOS (N) Y EL NÚMERO DE PARÁMETROS (P)

! Es conveniente no caer en la trampa de construir un modelo complejo (bastantes parámetros a estimar) con un número pequeño de datos!

En no pocas ocasiones, se encuentra un usuario muy feliz porque ha encontrado un modelo que tiene asociado un coeficiente de determinación muy alto, sin embargo al explorar con detenimiento se observa que con 10 datos ha construido un modelo

polinòmico de grado 8, lo cual es totalmente inconveniente. La razón es intuitivamente clara: si usted quiere ajustar una recta , con dos(2) datos, apriori, sin conocer cual es el problema y sin saber cuáles son los datos, podremos decir que el coeficiente de determinación será del 100%, pues sabemos que por dos puntos siempre pasa una recta. Lo mismo podremos decir de una parábola con tres (3) datos, y de un modelo de grado 8 con 8 datos.

Esto significa que el coeficiente de determinación no es confiable cuando la relación entre el número de datos con respecto al número de parámetros a estimar por mínimos cuadrados, es pequeña.

Regla empírica sobre la relación n/p. Como una guía empírica puede decirse que si existen aproximadamente 10 datos por cada parámetro que se desea estimar en el modelo, el valor del coeficiente de determinación que se calcule es confiable (creíble).

En general el coeficiente de determinación puede ajustarse de acuerdo con la relación del número de datos al número de parámetros, para encontrar el valor confiable del coeficiente de determinación, para un valor específico de n/p. Aquí se da origen al llamado **Coeficiente de Determinación Ajustado** (o corregido), el cual se presenta a continuación.

Coeficiente de Determinación Ajustado

Si se ha construido un modelo de regresión lineal que tiene **p** parámetros a estimar y se usaron en la estimación **n** datos, obteniendo un modelo con un coeficiente de determinación $\mathbf{R^2}$, el coeficiente de determinación ajustado $\mathbf{R_A^2}$ esta dado por :

$$R_A^2 \quad = \quad 1 \quad - \quad \frac{\sum e_i^2 / (n-p)}{\sum (y_i - \bar{y})^2 / (n-1)}$$

De donde resulta fácilmente que:

$$R_A^2 = 1 - \frac{n-1}{n-p}\left(1 - R^2\right)$$

En esta expresión se relaciona el coeficiente de determinación ajustado, con el ordinario. Veamos como funciona para algunos casos:

Ejemplo 1.

Si con 10 datos se construyera un polinomio de grado 8, el cual tiene nueve (9) parámetros y resultara con un coeficiente de determinación R^2= 90%, daría la falsa impresión de un buen modelo, sin embargo al calcular el Coeficiente de Determinación Ajustado resulta:

n=10 p=9 de donde:

$$R_A^2 = 1 - \frac{10-1}{10-9}\left(1 - 0.90\right) = 0.1$$

!!!! Tremenda Sorpresa !!!! Nos indica que en esas condiciones el valor creíble del coeficiente de determinación es el 10%.

Ejemplo 2

Supongamos la misma situación anterior pero donde lo único distinto es que todos los cálculos y estimaciones se realizaron con n=90 datos. Veamos que pasa:

$$R_A^2 = 1 - \frac{90-1}{90-9}\left(1 - 0.90\right) = 0.89$$

Paso de 90% a 89%, es decir que tuvo un cambio casi despreciable. Note que en esta ocasión se cumple la recomendación empírica de que hayan 10 datos por cada parámetro, es decir la razon n/p = 10.

EJERCICIOS PROPUESTOS

1. Resuelva el sistemas de ecuaciones (E 4.5) y muestre que para la familia de modelos rectilíneos $M(y/x) = ß_0 + ß_1 x$, el criterio de los mínimos cuadráticos concluye que:

$$\beta_1 = \frac{\frac{\sum x_i y_i}{n} - \overline{X}.\overline{Y}}{S_X^2}$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

2. Proporcione algún argumento intuitivo que permita convencerse de que el coeficiente de determinación de un modelo mínimo cuadrático para una familia polinómica es mayor, cuando más alto es el grado del polinomio a usar.

3. Pruebe que para la familia de modelos rectilíneos, es decir de la forma:

$$M(y/x) = ß_0 + ß_1 X$$

el coeficiente de determinación R^2 coincide con el cuadrado del coeficiente de correlación, r^2.

4. Plantee las ecuaciones normales, si en lugar de conocer los puntos (x_1, y_1), (x_2, y_2), ...,(x_n, y_n) sólo se conociera la distribución conjunta de frecuencias absolutas: $\{(x_i, y_i), n_{ij}\}$.

5. El "costo del mantenimiento" (Y) de cierto tipo de tractores parece incrementar con la "edad del tractor" (X). Con el propósito de encontrar un modelo que explique esta relación, se tomaron los siguientes registros:

Edad del tractor (X) (años)	Costo semestral del mantenimiento (Y) (en U.S)
4.5	619
1.0	549
4.0	495
4.5	1049
4.5	1033
5.0	1522
4.0	723
4.0	681
5.0	987
0.5	163
0.5	182
6.0	764
6.0	1373
1.0	978
1.0	466

a. Encuentre, para cada una de las siguientes familias, el modelo mínimo cuadrático.

5.1 $M(y/x) = \beta_0 + \beta_1 x$

5.2 $M(y/x) = \beta_0 + \beta_1 x^{1.5} + \beta_2 e^x$

5.3 $M(y/x) = \beta_0 x^{\beta_1}$

b. Cuál de los tres modelos encontrados le parece mejor, desde el punto de vista del ajuste. Use el coeficiente de determinación ajustado. Justifique.

c. El modelo mencionado en b. le parece bueno ? Comente.

d. Con base en el modelo encontrado en b. Haga la predicción para $X = 3.5$ años. Interprete muy claramente el valor obtenido.

6. A continuación se presentan los pesos iniciales (X) y aumentos de peso (Y) de 10 ratas hembras de 28 a 84 días de edad, sometidas a dieta de altas proteínas:

	Rata Número									
	1	2	3	4	5	6	7	8	9	10
Peso inicial X (gramos)	50	64	76	64	74	60	69	68	56	48
Aumento Y	128	159	158	119	133	112	96	126	132	118

a. Grafique el diagrama de dispersión.
b. Con base en el diagrama de dispersión proponga una familia de modelos para predecir el incremento de peso Y, con base en el peso inicial (X).

c. Por medio del criterio de los mínimos cuadrados, encuentre el mejor modelo de la familia propuesta que se ajusta a los puntos del diagrama de dispersión.

d. Comente sobre la bondad del modelo hallado.

e. Estime el aumento de peso promedio para las ratas con peso inicial de 70 grs.

7. A continuación se presentan registros sobre el precio (X) y la cantidad de naranja vendida en un supermercado, durante 12 días consecutivos.

Precio (X) $/ lbs.	100	90	80	70	70	70	70	65	60	60	55	50
Cantidad vendida lbs. (Y)	55	70	90	100	90	105	80	110	125	115	130	130

a. Ajuste un modelo rectilíneo para predecir la demanda (*Y*) con base en el precio.

b. Hágase una idea de la bondad del modelo a través del coeficiente de determinación.

c. Haga una estimación de la demanda cuando el precio por libra sea de $75, e interprete claramente el resultado.

8. Se piensa que la productividad en el trabajo de la construcción está relacionada con la duración del turno (jornada) de trabajo (en número de horas) por día. Para investigar el asunto se diseño un estudio. El cual arrojo los siguientes resultados, donde *x* es la duración la jornada en horas por día y *y* es la productividad (porcentaje de eficiencia).

(*x*,*y*)	No. de Observaciones
(6,50)	2
(6,70)	5
(6,90)	10
(8,50)	5
(8,70)	30

(*x*,*y*)	No. de Observaciones
(8,90)	25
(10,50)	8
(10,70)	25
(10,90)	11
(12,50)	10
(12,70)	6
(12,90)	2

a. Construya un diagrama de dispersión, graficando los puntos proporcionales del número de datos que representan.

b. De acuerdo con la forma del diagrama, plantee una familia razonable de modelos, para construir su modelo de regresión *M(y/x).*

c. Plantee las ecuaciones normales.

d. Estime e interprete los parámetros del modelo.

e. Encuentre en su modelo de regresión *M(y/x = 9),* interprete el resultado.

f. Qué jornada recomienda usted?

g. Dado que para cada valor de *x*, existen en los datos varios valores de *y*. Estime las varianzas: *V(y/x = 6), V(y/x = 8), V(y/x = 10), V(y/x = 12*). Le parece a usted que hay homogeneidad de varianzas?

h. Juzgue la bondad del modelo.

9. La siguiente tabla muestra datos de lluvias y filtraciones asociadas al rio Monocacy en Puente Jug, Maryland. (Tomado de Linsley and Franzini, 1964)

a. Con base en ellos construya un diagrama de dispersión y proponga algunos modelos que le parezcan plausibles para predecir la filtración media para un nivel dado de precipitación. *M(y/x).*

Lluvia No.	Precipitación *Y* (pulg.)	Filtración *X* (pulg.)
1	1.11	0.52
2	1.17	0.40
3	1.79	0.97
4	5.62	2.92
5	1.13	0.17
6	1.54	0.19
7	3.19	0.76
8	1.73	0.66
9	2.09	0.78
10	2.75	1.24
11	1.20	0.39
12	1.01	0.30
13	1.64	0.70
14	1.57	0.77
Lluvia No.	Precipitación *Y* (pulg.)	Filtración *X* (pulg.)

15	1.54	0.59
16	2.09	0.95
17	3.54	1.02
18	1.17	0.39
19	1.15	0.23
20	2.57	0.45
21	3.57	1.59
22	5.11	1.74
23	1.52	0.56
24	2.93	1.12
25	1.16	0.64

b. Ajuste por mínimos cuadrados los modelos propuestos por usted, y valore con base en el coeficiente de determinación ajustado.

c. Con base en el modelo que Ud. considero más adecuado haga la predicción correspondiente para una precipitación x = 2.3 pulg., interprete su respuesta en el contexto del problema

10. Un importante factor en la predicción de profundidad de escarcha para las vías pavimentadas es la temperatura media anual para el sitio en consideración.
La media de temperatura anual registrada en 10 diferentes estaciones meteorológicas en Virginia del Oeste son resumidos en la siguiente tabla.

Estación metereológica	elevación (pies)	latitud (grados)	temperatura media anual
Bayard	2375	39.27	47.5
Buckhannon	1459	39.00	52.3
Charleston	604	38.35	56.8
Flat Top	3242	37.58	48.4
Kearneysville	550	39.38	54.2
Madison	675	38.05	55.1
New Martinsville	635	39.65	54.4
Pickens	2727	38.66	48.8
Rainelle	2424	37.97	50.5
Wheeling	659	40.10	52.7

Puesto que un pavimento puede ser construido en distintos sitios de un estado donde los registros de temperatura no están disponibles, es necesario predecir la

temperatura media anual de la localidad con base en su elevación (altura sobre el nivel del mar) y latitud. Usando la información que aparece en la tabla realice:

a. La estimación por mínimos cuadrados de los parámetros $ß_0$, $ß_1$, $ß_2$ en el modelo $M(y/X_1\ X_2) = ß_0 + ß_1X_1 + ß_2X_2$ donde Y es la temperatura media anual (en grados Fahrenheit), X_1 la elevación en pies sobre el nivel del mar, X_2 latitud norte en grados.

b. Interprete claramente el significado de los valores obtenidos para $ß_0$, $ß_1$ y $ß_2$ en el contexto del problema.

c. Valore la importancia relativa de cada una de las 2 variables predictoras.

d. Calcule el coeficiente de determinación ajustado y juzgue la bondad de ajuste del modelo.

e. Use el modelo para realizar una predicción para $X_1 = 1000$ y $X_2 = 38°$ latitud. Interprete claramente su resultado.

11. La tabla a continuación se refiere al número de golpes N_i y su correspondiente fuerza de compresión libre de arcilla muy rígida q_i . Estime el coeficiente de correlación entre el número de golpes N_i y la fuerza de compresión q_i .

Número de golpes N_i	fuerza de compresión q_i
4	0.33
8	0.90
11	1.41
16	1.99
17	1.70
19	2.25
21	2.60
25	2.71
32	3.33
34	4.01
187	21.23

12. Se asume hipotéticamente que la concentración de sólidos disueltos y la turbidez de un arroyo son medidos simultáneamente por 5 días diferentes, seleccionados en forma aleatoria durante todo un año. Los datos son los siguientes.

día	sólidos disueltos	turbidez

	(mg/l)	(JTU)
1	400	5
2	550	30
3	700	32
4	800	58
5	500	20

Ya que la turbidez es fácil de medir se puede usar una ecuación de regresión para predecir la concentración de sólidos disueltos con base en la turbidez. Suponga que la varianza de concentración de sólidos es constante.

a. Ajuste una línea recta a estos datos. Que valores se obtuvo para el intercepto y la pendiente (parámetros de la recta de regresión).

b. Estime la desviación estándar de la concentración de sólidos disueltos a lo largo de la recta de regresión

c. Si no lo convence el modelo de línea recta, haga propuestas que le parezcan razonables.

13. Suponga que los datos del consumo de agua individual por día se acumularon para 4 barrios en Igor-City, tal como presenta la siguiente tabla.

a. Si el efecto del tamaño poblacional de un barrio, sobre el consumo individual es despreciable, determine la varianza muestral.

b. De los datos observados se nota una tendencia a creer en el consumo individual de agua con respecto al tamaño poblacional del barrio. Suponga que :

$$E(y/x) = \beta_0 + \beta_1 X$$

y que V(y/x) es constante para todo x.

i) Determine las estimaciones de mínimos cuadrados para β_0 y β_1

ii) Estime $S^2_{y/x}$

c. Un ingeniero está interesado en estudiar el consumo de agua en un población de 50.000 habitantes. Asuma distribución normal para Y. Determine la probabilidad de que la demanda de agua en la ciudad exceda 7 millones de galones diarios.

14. En la tabla a continuación se presenta la población de una comunidad para los años 1982 a 1992, que sugiere que la población en un año dado

depende de la población del año anterior, como predice el siguiente modelo: $X_t = a + bX_{t-1} + e$ donde X_t y X_{t-1} son los habitantes en el año t y $t - 1$, respectivamente, y e es un variable aleatoria normal con media 0 y desviación estándar σ.

Año	Habitantes
1982	240100
1983	245400
1984	247500
1985	251000
1986	253400
1987	258200
1988	261000
1989	262000
1990	265000
1991	268000
1992	274500

a. Con base a los datos de población dados, determine la estimación para a, b y σ.

b. Use el modelo y las estimaciones halladas para predecir la población para 1993.

REFERENCIAS BIBLIOGRAFICAS

BEHAR, R. **OJEDA M.M et AL.** (1996) “Un modelo de regresión jerárquica para análisis de datos de una muestra por conglomerados estratificada”. (Coautores: Ojeda M. M. Y Torres-Hernández M.T.). Ponencia presentada en el VI Encuentro de Estadísticos CUBA-MEXICO. Ciudad de La Habana, Cuba. 19-23 de Febrero de 1996

BEHAR, R. OJEDA M.M (1996) “A method for Predicting Means in Nonsampled Clusters using a Two-level Regression Model and a Computer Intensive Procedure ” Ponencia presentada en el VI Encuentro de Estadísticos CUBA-MEXICO. Ciudad de La Habana, Cuba. 19-23 de Febrero de 1996

BEHAR, R. OJEDA M.M (1995). La problemática de la enseñanza y el aprendizaje de la estadística en la educación superior. 1995). Sometido a la Revista "La Ciencia y el hombre" editada por la Universidad Veracruzana-México . ISSN 0187-8786..

BEHAR, R. OJEDA MM (1995). "Un replanteamiento del problema de la educación estadística: Perspectiva del aprendizaje" (Coautor: Doctor Mario Miguel Ojeda Ramírez- Universidad Veracruzana) 1995. Sometido a la Revista Heurística- Universidad del Valle ISSN 0121-262.

BEHAR, R., OJEDA MM , YEPES M. (1995). “Modelación del Rendimiento Escolar: Evaluación de los Criterios de Admisión de la Universidad del Valle Cali- Colombia” Ponencia presentada en la Conferencia Internacional CIMAF' 95, que sesionó bajo el lema: "Ciencia y Tecnología para el Desarrollo" durante los días 23 al 27 de Enero de 1995, durante los días 23 al 27 de Enero de 1995, La Habana-Cuba

BEHAR R., BOUZA C.N y OJEDA M.M. (1995). “Pruebas del tipo signo bajo modelos superpoblacionales”. Ponencia presentada en el II Congreso Iberoamericano de Estadística celebrado en septiembre de 1995 en OAXACA, México.

BEHAR, R. y YEPES M (1993). “Evaluación de los estimadores de mínimos cuadrados en un modelo linealizado cuando el término de error es aditivo”. Informe final de investigación. Octubre 1993. Presentado en las I Jornadas de Investigación en Ingeniería. Univalle, 1993

BEHAR, R. (1990). Métodos de validación de supuestos del Modelo de Regresión". Memorias I Simposio de Estadística de la Universidad Nacional, Bogotá, Mayo 31 - Junio 2 de 1990.

BEHAR, R. (1989). "Colinealidad: Peligrosa patología del modelo de regresión". Revista Heurística Vol 3 No. 1, Univalle. 1989.

BEHAR, R. (1987). . "Sobre el planteamiento del modelo de regresión". Revista Heurística Vol. 1, No.1. Univalle. 1987. ISSN -0121-262

BEHAR, R. (1983). "Bioestadística y Epidemiología ". Universidad del Valle, Facultad de Ingeniería.

BICKEL, P.J. and LEHMANN, EL (1975) " Descriptive statistics for non parametric models. Part 1 and 2 ". Annals of statistics 3: 1038 - 1044; 1045 - 1069. Part 1: Introduction. Part 2: Location.

BLALOCK, Hubert M, Jr. (Editor)(1974). " Measurement in the Social Sciences: Theories and strategies ". Chicago. Aldine.

BOX G.E.P. (1976). "Science and Statistics ". JASA 71: 791 - 799

BUHM D. y **D. Peat.** (1988). "Ciencia, orden y creatividad. Las raíces creativas de la ciencia y vida ". Editorial Kairos. Barcelona. Pp.299.

BUNGE M. (1986). " Intuición y razón ". Editorial Tecnos. Madrid. Pp.175.

CAMPBELL, N. (1953). " What is science? ". Dover publications Inc. New York. Pp.186.

CANSADO, E. (1958). "Estadística general ". Vol. 1 Cienes. Chile.

COCHRAM, W. G. (1965). "The Planning of observational studies of human populations ". Jour. Royal. Stat. Society A. Vol. 128: 234 - 265

COCHRAM, W. (1977). "Sampling techniques ". 3th. Ed. J. Wiley and Sons, N.Y.

CONOVER, W. J. (1980). "Practical nomparametric Statistics ". 2nd. Ed. N.Y.

COSTNER, Hebert L (1965) " Criteria for measures of association ". American Sociological Review 30: 341 - 353.

DAVID, F.N. (1962). " Games, goods and gambling: the origins and history of probability and statistical- ideas from the Earliest times to the Newtonians Era ". London Griffin; New York: Hafner.

DUNCAN, Otis, Dudley; CUZZORT, Ray P; and DUNCAN Beverly (1961). "Statistical geography : problems in analyzing areal data ". New York: Free Press.

ELASHOFF, Janet D. (1971). " Measures of association between a dichotomous and continuous variable. Page 218 - 230 in American statistical association social statistics section, proceedings. Washinton: the association.

FAIRLEY, William B; and MOSTELLER, Frederich (1977). "Statistics and public policy" Reading Mass: Addison - Wesley.

GOODMAN, Leo A. (1959). " Some alternatives ecological correlation ". American journal of Sociology 64: 610 - 625.

GOODMAN, Leo A. (1963). " On methods for: comparing contingency tables ". Journal of the royal statistical society series A 126: 94 - 108.

HACKING, Ian (1975). " The emergence of probability: A Philosophical study of early ideas about probability ". Cambridge Univ. Press.

JOINT COMMINTTEE OF THE CURRICULUM IN STATISTICS AND PROBABILITY OF THE AMERICAN STATISTICAL ASSOCIATION AND THE NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS (1973). " Reading, Mass: Addison Wesley.

Vol 1: "Exploring data ". Vol 2 : " Weighing chances ". Vol 3 : "Detecting patterns ". Vol 4: " Finding models ". Prepared and edited by a committee chaired by Frederick Mosteler and including William H. Kruskal; Richard Flink; Richards Pieters and Gerard R Rising.

KENDALL, Maurice G (1972). "Measurement in the study of society ".William Robson (Editor) " Man and the Social Sciences ". London: Allen and un win; Beverly Itills, Calif: Sage.

KENDALL, Maurice G, and PLACKETT R.L. (Editors) (1977) " Studies in the history of Statistics and probability. Vol 2. London : Griffin, New York: MacMillan.

KMENTA JAN, (1971). "Elements of Econometrics ". Ed. Collier MacMillan. New York.

KRUSKALL, William H. (1958). " Ordinal measures of association ". Journal of the American statistical association 53: 814 - 861.

KRUSKALL, William H. (1974). " The ubiquity of statistics ". American Statistician 28 No. 1: 3-6.

LARSEN, R., F. (1976). **"** Stroup Statistics in the real world ". Ed. Collier MacMillan. N.Y.

LEIK, Robert; and GOVE, Walter (1971). " Integrated approach to measuring association page 279 - 301 in Hebert L. Costner (Editor) Sociological methodology, 1971. San Francisco: Jossey-Bass.

LIEBERMAN, Bernhardt (1971). " Contemporany problems in statistics: A books of reading for the Behavioral sciences ". New York: Oxford Univ. Press.

LONGINO H. (1990). " Science as a social knowledge ". Princeton University Press. Pp.262.

MAISTOV, Leonid E. (1967). 1978 " Probability theory: A historical sketch ". Translated and edited by Samuel Kotz. New York: Academic Press.

MENDEZ, R.I.; NAMIHIRA, G.P.; MORENO, A.L. y SOSA DE M.C. (1984 1ª impresión. 1988 tercera reimpresión). " El protocolo de investigación ". " Lineamientos para su elaboración y análisis ". Trillas, México. Pp.210.

MENDEZ, R.I. (1986a). " Causalidad en medicina ". Gaceta médica de México. Vol. 122, Nos. 1 y 2.

MENDEZ, R.I. (1986b). " La Estadística y la Epidemiología. Niveles de estudio de la ciencia y aspectos históricos ". Revista de la Facultad de Medicina UNAM 29(6): Pp.261 - 266.

MENDEZ, R.I. (1988). " La estructura de la investigación estadística ". Comunicaciones técnicas IIMAS - UNAM. Serie azul. No. 106.

MENDEZ, R.I. (1989a). " La ubicación de la Estadística en la metodología científica ". Ciencia 40: 39 - 48.

MEANS, R. K. (mayo 1965). "Interpreting statistics: and avt, nursing outlook ". 13 Pp.34 - 37.

MENDENHALL, W. (1979). " Introducción a la probabilidad y a la Estadística ". Ed. Wadsworth international Iberoamérica.

MOSTELLER, Frederick and TUKEY, John W. (1977) " Data analysis and regression : A second course in statistics " Reading Mass : Addison Wesley.

MOSTELLER, KRUSKALL, TANNUR. (1980). " Statistics, a guide to unknown. Holdenday ".

NETER, John and Maynes E. Scott (1970). " On the appropriateness of the correlation coefficient with a 0 - 1 dependent variable. JASA, 65: 501 - 509.

NIETO DE ALBA, U. (1978). " Introducción a la Estadística. Concepción clásica y bayesiana ". Vol. 1 Editorial Aguilar.

NIETO DE ALBA, Ubaldo (1974). "Introducción a la estadística descriptiva ". Editorial Aguilar. Madrid, España .

PEARSON, Karl. (1921-1933) 1978. "The history of the statistics in the seventeenth and eighteenth centuries. Against the changing background of intellectual, scientific and religions thorght" Edited by E.S. Pearson. London: Griffin; New York: MacMillan.

PEARSON E.S. and KENDALL, Maurice. (Editors) (1970). " Studies in the history of statístics and probability" Vol.1 London: Griffin: Darien Com: Hafner.

SCHMID, C.; SCHMID, S. (1979). " Hand book graphic presentation. J. Wiley and Sons "

STURGES, Herbert A. (1926) " The choice of a class interval ". American Statistical Association, Vol.21, Pp. 65 - 66.

TANUR, Judith M. et al (Editors) (1972). " Statistics: a guide to the unknown". San Francisco: Holden - day.

THELL, H. (1971). " Principles of Econometrics ". Ed. John Wiley and Sons, N.Y.

TODHUNTER, Isaac. (1965) 1949. "A history of the mathematical Theory of probability from the time of Pascal to that of Laplace ". New York: Chelseca.

TUKEY, John W. (1970). 1977. "Exploratory data analysis" Reading Mass: Addison Wesley.

UNGER LEIDER H, SMITH C.C., (Febrero 1967) " Use and abuse of Statistics, geriatrics", 22, 112 - 120.

WALKER, Helen. (1929). " Studies in the history of statístical method, with special reference to artain educational problems". Baltimore: Williams and Wilkins.

WESOLOWSKY, G. (1979). " Multiple regression and analysis of variance ". Wiley and Sons, N.Y.

WESTERGAARD, Harold. (1932). " Contributions to the history of Statistics ". London: King.

YULLE-KENDALL. (1967). "Introducción a la estadística matemática". Ed. Aguilar, Madrid.

ZINMERMAN, J. P. (1969). " Statistical data and their use. Physical therapy ", 49 301 - 302.